sensors

MDPI

Article

CNN-Based Multimodal Human Recognition in Surveillance Environments

Ja Hyung Koo, Se Woon Cho, Na Rae Baek, Min Cheol Kim and Kang Ryoung Park *

Division of Electronics and Electrical Engineering, Dongguk University, 30 Pil-dong-ro, 1-gil, Jung-gu, Seoul 100-715, Korea; koo6190@dongguk.edu (J.H.K.); jsu319@dongguk.edu (S.W.C.); naris27@dongguk.edu (N.R.B.); mincheolkim@dongguk.edu (M.C.K.)

* Correspondence: parkgr@dongguk.edu; Tel.: +82-10-3111-7022; Fax: +82-2-2277-8735

Received: 7 August 2018; Accepted: 8 September 2018; Published: 11 September 2018

Abstract: In the current field of human recognition, most of the research being performed currently is focused on re-identification of different body images taken by several cameras in an outdoor environment. On the other hand, there is almost no research being performed on indoor human recognition. Previous research on indoor recognition has mainly focused on face recognition because the camera is usually closer to a person in an indoor environment than an outdoor environment. However, due to the nature of indoor surveillance cameras, which are installed near the ceiling and capture images from above in a downward direction, people do not look directly at the cameras in most cases. Thus, it is often difficult to capture front face images, and when this is the case, facial recognition accuracy is greatly reduced. To overcome this problem, we can consider using the face and body for human recognition. However, when images are captured by indoor cameras rather than outdoor cameras, in many cases only part of the target body is included in the camera viewing angle and only part of the body is captured, which reduces the accuracy of human recognition. To address all of these problems, this paper proposes a multimodal human recognition method that uses both the face and body and is based on a deep convolutional neural network (CNN). Specifically, to solve the problem of not capturing part of the body, the results of recognizing the face and body through separate CNNs of VGG Face-16 and ResNet-50 are combined based on the score-level fusion by Weighted Sum rule to improve recognition performance. The results of experiments conducted using the custom-made Dongguk face and body database (DFB-DB1) and the open ChokePoint database demonstrate that the method proposed in this study achieves high recognition accuracy (the equal error rates of 1.52% and 0.58%, respectively) in comparison to face or body single modality-based recognition and other methods used in previous studies.

Keywords: multimodal human recognition; surveillance environment; CNN; human recognition by face and body

1. Introduction

Previous biometrics studies have used various modalities, including the face, fingerprints, body, irises, retinas veins, and voice [1–9]. In a typical surveillance camera environment, it is difficult to use fingerprints or vein recognition, so face, body, and iris methods have been considered. In the case of iris recognition, a zoom lens and a near-infrared (NIR) light illuminator of high power are needed to capture iris images at a distance, so the systems are large and expensive and can be used in a limited range of contexts. Also, in a surveillance environment, the camera is normally installed above the user and captures images in a downward direction, so it mainly takes off-angle images that capture the user's iris at an angle. In such circumstances, the recognition accuracy is greatly reduced [9].

Face recognition has often been considered for surveillance environments as it can generally be conducted in a visible light camera environment. However, in a surveillance environment, most cases involve a camera capturing images in a downward direction from above and people do not look directly at the camera. Thus, it is generally difficult to capture front facial images, and in such cases, facial recognition accuracy is greatly reduced. To address this issue, 105 composite geometrical descriptors for 3D face analysis based on 3D face data captured by a laser scanner were presented in a previous study [10]. The authors mapped these new descriptors on 217 facial depth maps and analysed them based on descriptiveness of facial shape and exploitability for detecting landmark points. In other research [11], they proposed a method to automatically diagnose and formalize prenatal cleft lip with key points and recognize the type of defect in 3D ultrasonography. For that, they adopted differential geometry as a framework to describe facial curvatures and shapes. In previous research [12], Cowie, et al. introduced the various methods of emotion recognition in human-computer interaction including applications, framework, input and output-related issues, physiological and domain issues, training and test materials, and case study, etc. In [13], Tsapatsoulis et al. proposed the method of face extraction from non-uniform background based on the fusion of a retrainable neural network and morphological size distribution method. In addition, they also proposed the face recognition in MPEG-4 compressed domain to fuse the face images of high quality and low computational complexity.

In this research, we consider using face and body data for human recognition in a visible light camera surveillance environment, which is based on the movements of people's bodies and the texture, color, and shape of their bodies. However, when images are captured by a camera installed in an indoor surveillance environment, in many cases, part of the target body is not included in the camera viewing angle, and only part of the body is captured, which causes a drop in human recognition accuracy. Aside from this, a study was conducted on human recognition using body images captured by visible light and thermal cameras [14], but this method requires high-cost thermal cameras, so it is not suitable for use in a normal surveillance environment. The next section analyzes previous studies on human recognition using face and body data in a surveillance camera environment.

2. Related Work

Previous studies on human recognition in a surveillance environment using face and body data can be broadly divided into single-modality-based methods and multiple-modality-based methods. The former include face recognition, movement-based body recognition, as well as texture-, color-, and shape-based body recognition. In a study on face recognition in a surveillance camera environment, Kamgar-Parsi et al. detected face regions through the boosted classifier of Haar wavelets method and performed morphing of facial images based on an active shape model (ASM), followed by chi-square distribution-based classification [15]. An et al. used several cameras to capture face images and performed face recognition based on a dynamic Bayesian network (DBN) [16]. Grgic et al. installed five cameras above a door and captured face data at three set locations to perform face recognition using a principal component analysis (PCA) method [17]. Banerjee et al. performed recognition using a soft-margin-based learning method for multiple feature-kernel combinations (SML-MKFC) with domain adaptation (DA) [18]. In this type of surveillance environment, it is often difficult to capture a front face image, hence, the face recognition accuracy is reduced. To address this problem, there have been studies [19] on recognition by extracting landmark points on a face and using these to adjust the face to a front angle (face frontalization). However, due to the nature of surveillance camera environments (especially indoor environments), in many cases, the face region is subject to optical and motion blurring due to the target moving at a short distance from the camera. In such cases, facial landmark point extraction is not precise, so face frontalization cannot be performed. Therefore, recognition methods have been developed that use the body region's texture, color, and shape information obtained from a single image, as well as methods that use body motion from several continuous images. Methods of the former kind include the following.

Though not focused on human identification, Antipov et al. compared the performance of hand-crafted feature-based person re-identification and a histogram of oriented gradients with the performance of learned features that were based on a mini-convolutional neural network (CNN) and AlexNet-CNN for the purpose of gender recognition [20]. Layne et al. studied body image-based recognition using a method of symmetry-driven accumulation of local features (SDALF) with metric learning attributes (MLA) [21]. Nguyen et al. performed a gender recognition study using histogram of oriented gradient (HOG), PCA, and support vector machine (SVM) on user body images captured by visible light and thermal cameras [22]. Also, in [14], AlexNet and PCA-based feature extraction and distance measurement were used to perform a study on personal identification based on user body images captured by visible light and thermal cameras. Figueira et al. performed a study on person re-identification based on a semi-supervised multi-feature learning (MFL) method [23]. In [24], a method of person re-identification was proposed that uses spatial covariance regions of human body parts and spatial pyramid matching. Prosser et al. used a Gabor and Schmid filter to perform feature extraction and then used ensemble ranking SVM to perform person re-identification [25]. Ensemble ranking SVM was proposed as a method to overcome the scalability limitation of existing SVM-based ranking problems. Chen et al. used a spatially constrained similarity function on a polynomial feature map (SCSP) and PCA to perform feature extraction and performed a study on person re-identification based on spatial pyramid matching [26]. Liao et al. performed a study on person re-identification based on local maximal occurrence (LOMO) and cross-view quadratic discriminant analysis (XQDA) [27]. In [28], person re-identification was performed using a filter pairing neural network (FPNN) to resolve the problems of misalignment, photometric and geometric transforms, occlusions, and background clutter. In both [29,30] a Siamese CNN (S-CNN) structure was used for person re-identification, but the methods were different in that [29] used 7 convolution blocks, whereas [30] used two convolutional layers. In [31], a positive mining method was proposed for training a CNN for person re-identification, and discriminative deep metric learning (DDML) was applied. Yang et al. proposed training multi-level (e.g., pixel-level, patch-level, and image-level) descriptors using weighted linear coding (WLC) for person re-identification [32].

In these ways, recognition methods that use texture, color, and shape data from body regions obtained from a single image can compensate for their face recognition disadvantages, but they still have the disadvantage that they can misidentify an imposter as being the genuine person if they wear clothes of a similar color or arrangement. They can also show degraded recognition performance when part of the target body is not captured in an image. To resolve this problem, recognition methods have been proposed that use body motion and so forth in several continuous images. In [33], a study was performed on PCA and silhouette analysis-based gait recognition for human identification. In [34], gait-based recognition was further investigated, but to resolve the problem of insufficient existing gait data, synthetic gait energy images (GEI) were obtained, and the synthetic GEI and the features extracted from PCA and multiple discriminant analysis (MDA) were combined and was applied to improve recognition performance. However, these studies have the disadvantage that they can mainly be used with continuous images of a person's side (images in which the person is moving perpendicular to the direction from which the camera is shooting), but it is difficult to use them when the person approaches or moves further away from the camera.

In view of these problems, multimodal human recognition methods that combine face and body data have been proposed, and these previous methods mainly combine side face recognition and movement-based body recognition. In [35,36], feature extraction was performed using enhanced side-face images (ESFI) and the GEI method. Then, PCA and MDA were performed and recognition was done through a fusion method based on sum, product, and max rules. In [37], side face and GEI features were converted using PCA and MDA, respectively, and combined at the feature level to perform recognition. In [38], side face recognition was done through curvature-based matching, and gait recognition was done through direct GEI matching. Then the two types of recognition data were combined through a sum rule and a product rule. Kale et al. used posterior distribution and template matching methods for gait and side face recognition, and they performed fusion through sum, min, and product rules [39].

Table 1. Summary of our study and previous works on human recognition.

Category		Method	Advantage	Disadvantage
Single modality-based	Face recognition	ASM and image morphing [15]	Not affected by changes in people's clothes, etc. In comparison to body recognition, few cases occur where part of the region is not captured or pose variation happens.	Difficult to capture front face images. Face frontalization is difficult due to motion and optical blurring in the captured face images.
		DBN [16]		
		PCA [17]		
		SML-MKFC with DA [18]		
		ResNet [40,41]		
	Texture-, color-, and shape-based body recognition using single frame	AlexNet-CNN, HOG, and Mini-CNN[20], VGG [42,43]	Using body information, which has a larger area than the face, and recognition at long distances is possible.	Can misidentify an imposter as being the genuine person if they wear the same clothes. Reduced recognition performance in case that part of the target body is not captured.
		SDALF + MLA [21]		
		CNN + PCA [14]		
		HOG + PCA + SVM [22]		
		Semi-supervised MFL [23]		
		Spatial covariance region [24]		
		SCSP + SPM [26]		
		FPNN [28]		
		S-CNN [29,30]		
		CNN + DDML [31]		
		Multi-level descriptor by WLC [32]		
		LOMO + XQDA [27]		
		Ensemble ranking SVM [25]		
	Body movement (gait)-based recognition using multiple frames	PCA + silhouette analysis-based gait recognition [33]	Higher recognition accuracy than body recognition based on a single image.	Difficult to use when a person approaches or moves further away from the camera. By processing continuous images, the processing time is long.
		Synthetic GEI, PCA + MDA [34]		
Multiple modality-based	Side face recognition + body movement (gait)-based recognition using multiple frames	ESFI + GEI [35,36]	Higher recognition accuracy than single modality-based methods for face recognition or body movement-based recognition.	
		Side face + GEI [37]		
		Curvature-based matching + direct GEI [38]		
		Posterior distribution + template matching [39]		
		Image-based VH [44]		
		View-normalized sequences [45]		
		KFA + RSM framework [46]		
		Eigenface calculation +α-GEI [47]		
		HMM + Gabor-based EBGM [48]		
		Fisherface + silhouette image-based LPP [49]		
	Frontal face and texture-, color-, and shape-based body recognition using single frame	MLBP + PCA [50,51], HOG [52]	– Higher recognition accuracy than single modality-based methods – By single image processing, processing speed is fast.	Lower accuracy than deep CNN-Based method
		Deep CNN-based multimodal human recognition using both face and body (Proposed method)		Requiring an intensive training process of CNN

In [44], a study on face and gait recognition using image-based visual hull (VH) was performed. In [45], view-normalized sequences were used to perform gait recognition and face recognition, and they were combined through a cross-modal fusion rule to improve recognition performance. Guan et al. performed a study in which face recognition based on kernel Fisher analysis (KFA) was combined with gait recognition based on a random subspace method (RSM) [46]. Hofmann et al. used the eigenface calculation and α-GEI methods for a combined recognition of face and gait, respectively, in the Human ID Gait Challenge [47]. Liu et al. performed a study on the combined recognition of face and gait based on a hidden Markov model (HMM) and Gabor feature-based elastic bunch graph matching (EBGM) methods [48]. Also, Geng et al. performed a study on distance-driven fusion of face recognition, which was based on Fisherface, and gait recognition, which was based on silhouette image-based locality preserving projection (LPP) [49]. Most of these methods were applied to continuous images of a person's side (images in which the person is moving perpendicularly to the camera's shooting direction), and these methods have the disadvantage of being difficult to apply when the person is approaching or moving further away from the camera. In addition, they must process several continuous images, so they also have the drawback of a long processing time. To resolve these problems, this paper presents a deep CNN-based multimodal human recognition method that uses both face and body data in a single image. In addition, this method can be used to perform recognition in cases where the person is approaching or moving further away from the camera. These cases occur frequently in indoor surveillance (especially hallway) environments, but they were not sufficiently addressed by previous studies. Table 1 shows the advantages and disadvantages of the methods proposed in previous studies on human recognition in a surveillance camera environment and present study.

3. Contribution of Our Research

Our research is novel in the following four ways in comparison to previous works:

- Previous methods for face- and body-based multimodal human recognition have mainly been based on continuous images of the side face and gait captured during lateral movement relative to the camera. However, this study focuses on cases that often occur in indoor surveillance camera environments (especially hallways) in which a person is approaching or moving further away from the camera; the proposed method is the first approach for the multimodal human recognition that separately recognizes face and body regions in a single image and combines them.
- The person's whole body image is not used as a single CNN input. Rather, the face region and the body region are separated, and each is used as a separate CNN input. Thus, more detailed texture, color, and shape information regarding each region can be used. As a result, the recognition accuracy can be improved beyond that of methods that use whole body images as a single CNN input.
- A visual geometry group (VGG) Face-16 CNN is used for the face region, and a residual network (ResNet)-50 CNN is used for the body region. The body region is larger than the face region, and more detailed texture, color, and shape data must be extracted from the clothes and body. Therefore, the ResNet-50 is used because it has more layers and uses detailed residual information. On the other hand, the face region is smaller than the body region, and recognition normally uses more mid- or low-frequency information than high-frequency information, so the VGG Face-16 is used rather than the ResNet-50, which uses detailed residual information.
- Unlike previous methods that only focus on cases in which the entire body is included in the input image, the targets of the proposed method also include images in which part of the body region cannot be seen in the input image. To make impartial comparison experiments possible, the Dongguk face and body database (DFB-DB1), which was custom made using two kinds of cameras to evaluate performance in a variety of camera environments, and the VGG Face-16 and ResNet-50 CNN models were made public to other researchers in [53].

4. Proposed Method

4.1. Overall Procedure of Proposed Method

Figure 1 shows an overall flowchart of the proposed method. First, the face region in an image captured by a surveillance camera is detected by the adaptive boosting (AdaBoost) detector [54]. Then, a more accurate face region is detected based on the positions of the facial features (both eyes) detected by the dlib facial feature tracker [55] (step (1) in Figure 1). After this, the body region is defined based on the position and size of the detected face region (step (2) in Figure 1). In the next step, the face region's focus score is measured and the next recognition step is only performed if this value is above a certain threshold (steps (3) and (4) in Figure 1). If it is not, the next image is acquired from the camera. After this, the CNN models are run using the face region and body region as separate inputs (steps (5) and (6) in Figure 1). The extracted CNN features are used to measure their distance from the already registered features (steps (7) and (8) in Figure 1). Score-level fusion is performed using the two obtained distances, and a final matching score is obtained. This is then used to perform human recognition (steps (9) and (10) in Figure 1).

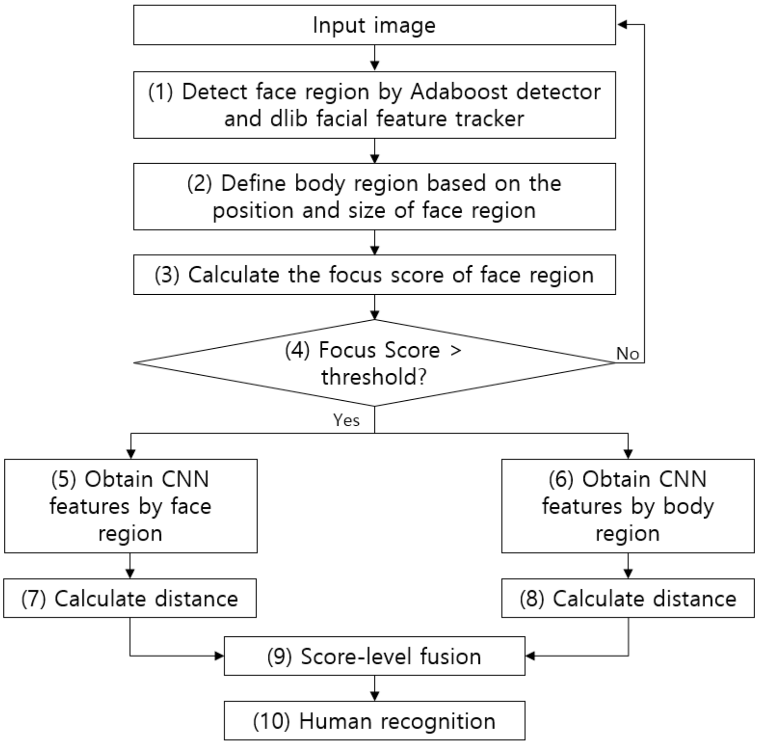

Input image
(1) Detect face region by Adaboost detector and dlib facial feature tracker
(2) Define body region based on the position and size of face region
(3) Calculate the focus score of face region
(4) Focus Score > threshold?
No
Yes
(5) Obtain CNN features by face region
(6) Obtain CNN features by body region
(7) Calculate distance
(8) Calculate distance
(9) Score-level fusion
(10) Human recognition

Figure 1. Overall procedure of proposed method.

4.2. Detection of Face and Body Regions as well as Focus Measurement

As explained in Section 4.1 and shown in Figure 2, the AdaBoost detector is used to detect the face region in an image captured by a camera [54]. AdaBoost detector uses the cascaded weak classifiers based on Haar feature, and it has been widely used for face detection. In this research, we used the AdaBoost detector provided from OpenCV library [56] without additional training with our experimental images. AdaBoost detector can generate the roughly detected face box which includes face and the part of background. Therefore, a more accurate face region is detected based on the positions of facial features (both eyes) detected by the dlib facial feature tracker [55]. In this research, we used the open source of dlib facial feature tracker provided from [55] without additional training with our experimental images. Also, as explained in Section 4.1, the body region is defined based on the size and position of the detected face region and anthropometric data on a normal person's body, as shown in Figure 2d. In details, based on the center position ($x_{\_face}$, $y_{\_face}$), width ($w_{\_face}$), and height ($h_{\_face}$) of the detected face region, the center position ($x_{\_face}$, $y_{\_face} + 1.8 \times h_{\_face}$), width ($1.8 \times w_{\_face}$), and height ($2.2 \times h_{\_face}$) of body box are defined, respectively. The lowest vertical position of body box is limited by "image height—1". In addition, the left- and right-most positions of body box are limited by "0" and "image width—1", respectively.

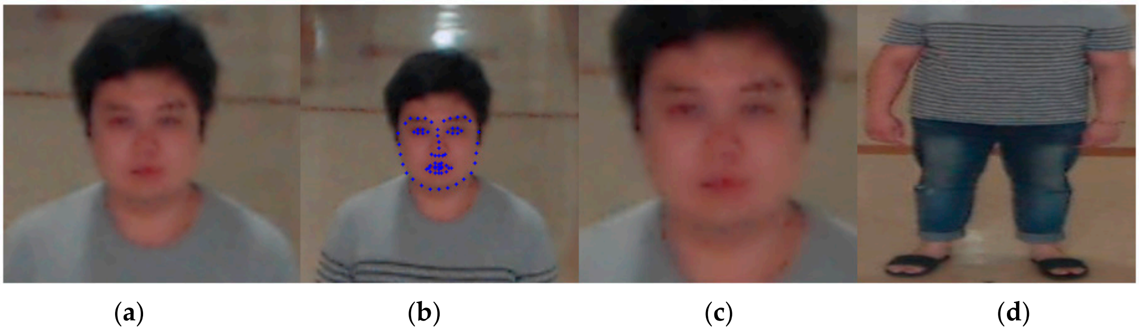
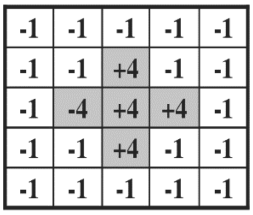


Figure 2. Detection of face and body region. (**a**) Face region detected by AdaBoost algorithm from input image, (**b**) facial landmarks detected in face region by dlib facial feature tracker, (**c**) redefined face region based on eye landmarks, (**d**) defined body region.

After this, the 5 × 5 mask proposed in [57] is used on the face region to calculate the focus score. The shape of this mask is shown in Figure 3. The 5 × 5 mask was designed to measure the amount of high frequency component in image [57]. In details, the magnitude value is computed by the convolution operation with the 5 × 5 convolution kernel in the image based on the moving step of 1 pixel both in horizontal and vertical directions as shown in Equations (1) and (2). Then, this magnitude value (FS of Equation (2)) is normalized so as to be presented in the range from 0 to 100 based on min-max scaling, and min and max values were determined from the training data. This normalized value is used as final focus score, and the higher the score, the better the focus condition. The next recognition step is only performed if this focus score is above a certain threshold. If it is not, the next image is acquired from the camera instead of recognition. The optimal threshold of focus score was experimentally determined as 20 (both in DFB-DB1 and ChokePoint databases) from the training data so as to obtain the highest accuracy of recognition.



-1	-1	-1	-1	-1
-1	-1	+4	-1	-1
-1	-4	+4	+4	-1
-1	-1	+4	-1	-1
-1	-1	-1	-1	-1

Figure 3. The 5 × 5 mask for focus assessment.

$$O[x,y] = I[x,y]M[x,y] = \sum_{q=0}^{H-1} \sum_{p=0}^{W-1} I[p,q]\, M[x-p, y-q] \tag{1}$$

$$\mathrm{FS} = \left(\sum_{y=0}^{H-1} \sum_{x=0}^{W-1} O[x,y]\right) / (W \times H) \tag{2}$$

In Equations (1) and (2), $I[x,y]$, $O[x,y]$ and $M[x,y]$ are input, output, and 5 × 5 mask images, respectively. W and H are the image width and height, respectively. In the DFB-DB1 database, which was custom made for this study, images were captured by two types of cameras, namely, the Logitech BCC950 [58] and the Logitech C920 [59], to evaluate performance of the proposed method in a variety of camera environments. Figure 4 shows the focus scores of images in DFB-DB1. Also, Figure 5 shows the focus scores of the ChokePoint dataset [60], which is an open database used in this study. As seen by a comparison of Figures 4b and 5b, the blurring due to user movement is more severe in the images in Figure 4b.

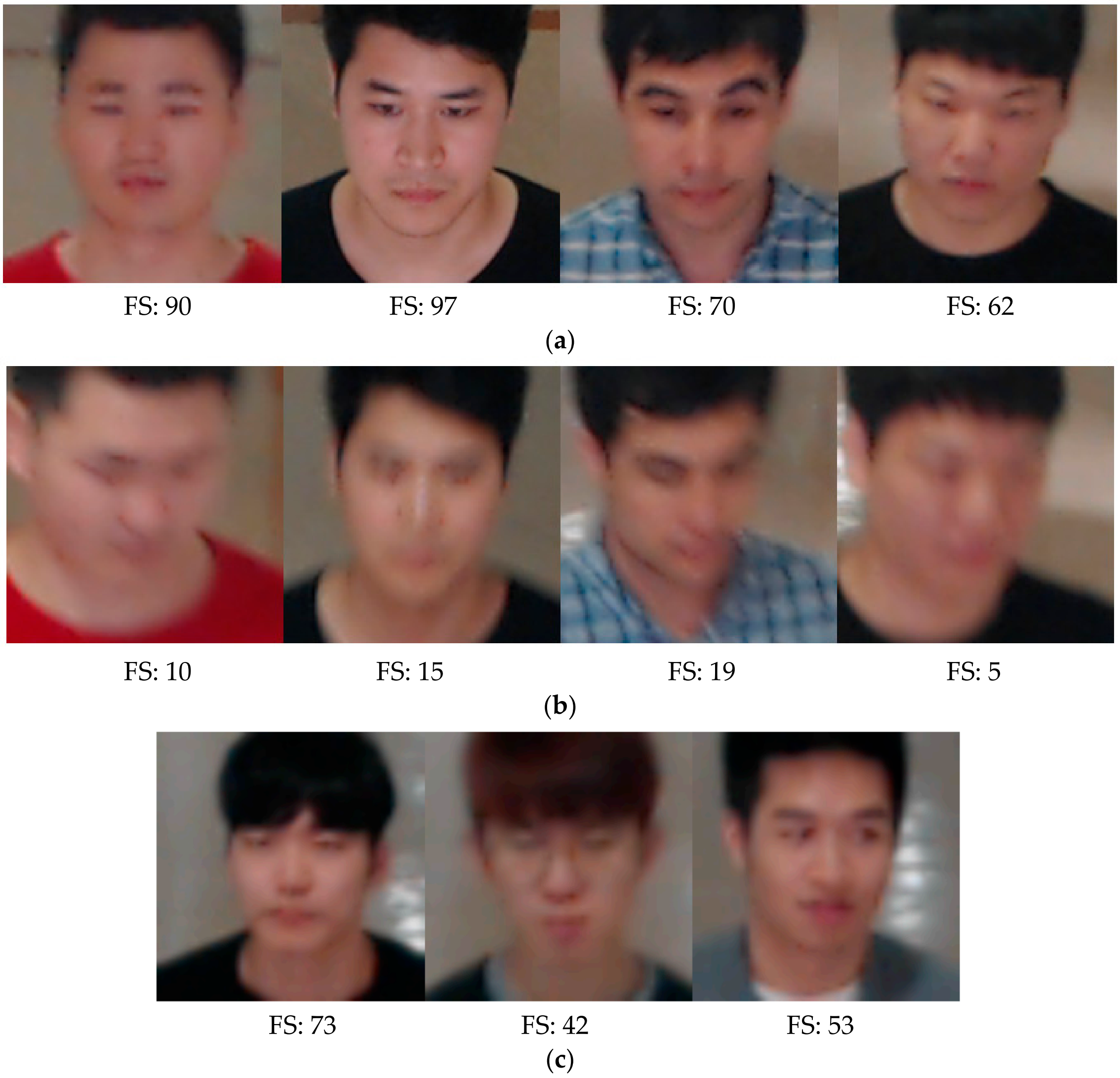


Figure 4. DFB-DB focus score calculation example images: (**a**) Examples of images with a focus score (FS) above the threshold (20) (BCC950 camera images). (**b**) Examples of images with a focus score below the threshold (BCC950 camera images). (**c**) Examples of images with a focus score above the threshold (20) (C920 camera images).

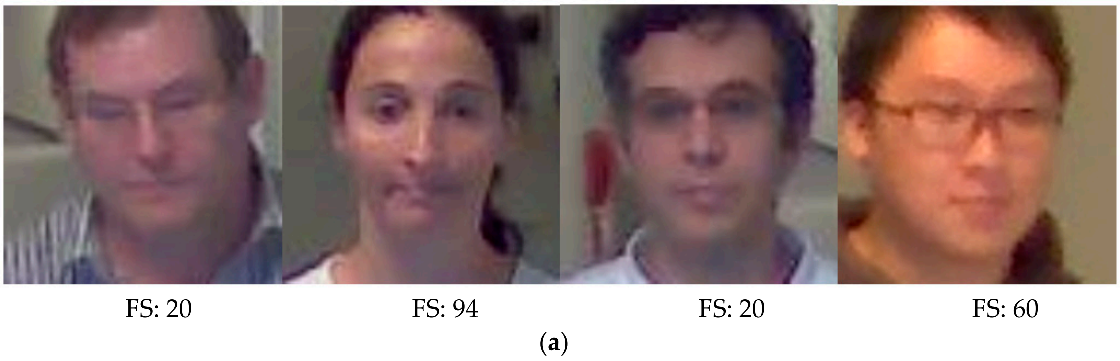
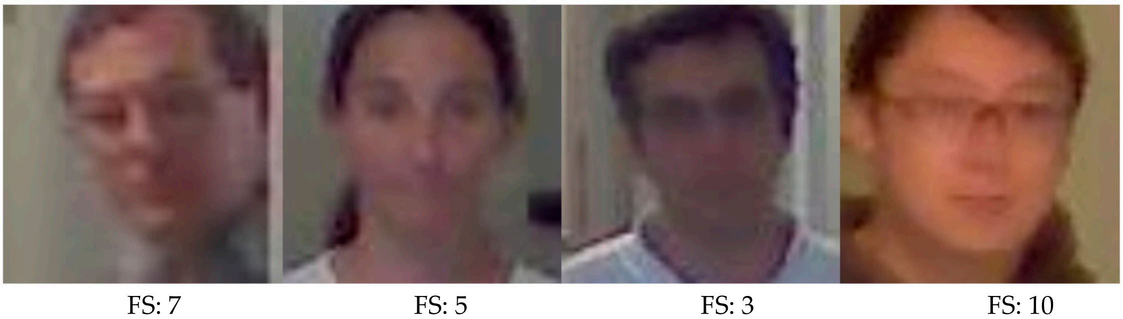


Figure 5. *Cont.*



(**b**)

Figure 5. ChokePoint dataset focus score calculation example images: (**a**) Examples of images with a focus score (FS) above the threshold (20). (**b**) Examples of images with a focus score below the threshold.

4.3. CNN for Face Recognition

In the proposed method, face recognition is performed using the VGG Face-16 CNN model, which takes the facial regions obtained in Section 4.2 as input. The VGG Face-16 CNN model is used for the face region, and the ResNet-50 CNN model is used for the body region. We used the VGG Face-16 CNN model provided from [61] in this research. The body region is larger than the face region, and more detailed texture, color, and shape data must be extracted from the clothes and body, so the ResNet-50 is used, as it has more layers and detailed residual information. On the other hand, the face region is smaller than the body region, and recognition normally uses more mid- or low-frequency information than high-frequency information, so the VGG Face-16 is used rather than the ResNet-50, which uses detailed residual information. To fine-tune the pre-trained VGG Face-16 model [3] with the database used in this study, the detected face regions from Section 4.2 are normalized to a 224 × 224 pixels size. The normalization was performed by bi-linear interpolation. VGG Face-16 has the same structure as VGG Net-16 with 13 convolutional layers, 5 pooling layers, and 3 fully connected layers, as shown in Figure 6 and Table 2. VGG Face-16 and VGG Net-16 have no structural differences, but they were trained differently. That is, VGG Face-16 is a model trained with labeled faces in the wild [62] and YouTube faces [63], while VGG Net-16 [64] is a model trained in the ImageNet large-scale visual recognition competition (ILSVRC)-2014 [65]. Normally, the size of a feature map obtained from the convolution operation in a CNN is calculated from the width or height of the filter, the width or height of the input image (or feature map) before it enters the convolutional layer, the amount of padding in the convolutional layer, and the number of strides [66]. After passing through the convolution layer, the rectified linear unit (ReLU) layer [67] is next. Normally, non-overlapping pooling windows obtain better results [68], so a filter size of 2 × 2 with a stride of 2 × 2 was used in this study. The final layer is the fully connected layer (FCL). In the 3rd FCL, there is a softmax layer. Finally, to avoid overfitting in the training data used during fine-tuning, dropout layers are used in the 1st and 2nd FCLs. In this study, the dropout layer probability was set at 50%.

Table 2. Descriptions of VGG Face-16 model.

Layer Type		Number of Filters	Size of Feature Map	Size of Filter	Number of Strides	Amount of Padding
Image input layer			224 (height) × 224 (width) × 3 (channel)			
Group 1	Conv1_1 (1st convolutional layer)	64	224 × 224 × 64	3 × 3	1 × 1	1 × 1
	ReLU1_1		224 × 224 × 64			
	Conv1_2 (2nd convolutional layer)	64	224 × 224 × 64	3 × 3	1 × 1	1 × 1
	ReLU1_2		224 × 224 × 64			
	MaxPool1	1	112 × 112 × 64	2 × 2	2 × 2	0 × 0
Group 2	Conv2_1 (3rd convolutional layer)	128	112 × 112 × 128	3 × 3	1 × 1	1 × 1
	ReLU2_1		112 × 112 × 128			
	Conv2_2 (4th convolutional layer)	128	112 × 112 × 128	3 × 3	1 × 1	1 × 1
	ReLU2_2		112 × 112 × 128			
	MaxPool2	1	56 × 56 × 128	2 × 2	2 × 2	0 × 0
Group 3	Conv3_1 (5th convolutional layer)	256	56 × 56 × 256	3 × 3	1 × 1	1 × 1
	ReLU3_1		56 × 56 × 256			
	Conv3_2 (6th convolutional layer)	256	56 × 56 × 256	3 × 3	1 × 1	1 × 1
	ReLU3_2		56 × 56 × 256			
	Conv3_3 (7th convolutional layer)	256	56 × 56 × 256	3 × 3	1 × 1	1 × 1
	ReLU3_3		56 × 56 × 256			
	MaxPool3	1	28 × 28 × 256	2 × 2	2 × 2	0 × 0
Group 4	Conv4_1 (8th convolutional layer)	512	28 × 28 × 512	3 × 3	1 × 1	1 × 1
	ReLU4_1		28 × 28 × 512			
	Conv4_2 (9th convolutional layer)	512	28 × 28 × 512	3 × 3	1 × 1	1 × 1
	ReLU4_2		28 × 28 × 512			
	Conv4_3 (10th convolutional layer)	512	28 × 28 × 512	3 × 3	1 × 1	1 × 1
	ReLU4_3		28 × 28 × 512			
	MaxPool4	1	14 × 14 × 512	2 × 2	2 × 2	0 × 0
Group 5	Conv5_1 (11th convolutional layer)	512	14 × 14 × 512	3 × 3	1 × 1	1 × 1
	ReLU5_1		14 × 14 × 512			
	Conv5_2 (12th convolutional layer)	512	14 × 14 × 512	3 × 3	1 × 1	1 × 1
	ReLU5_2		14 × 14 × 512			
	Conv5_3 (13th convolutional layer)	512	14 × 14 × 512	3 × 3	1 × 1	1 × 1
	ReLU5_3		14 × 14 × 512			
	MaxPool5	1	7 × 7 × 512	2 × 2	2 × 2	0 × 0
Fc6 (1st fully connected layer)			4096 × 1			
ReLU6			4096 × 1			
Dropout6			4096 × 1			
Fc7 (2nd fully connected layer)			4096 × 1			
ReLU7			4096 × 1			
Dropout7			4096 × 1			
Fc8(3rd fully connected layer)			#classes			
Softmax layer			#classes			
Output layer			#classes			

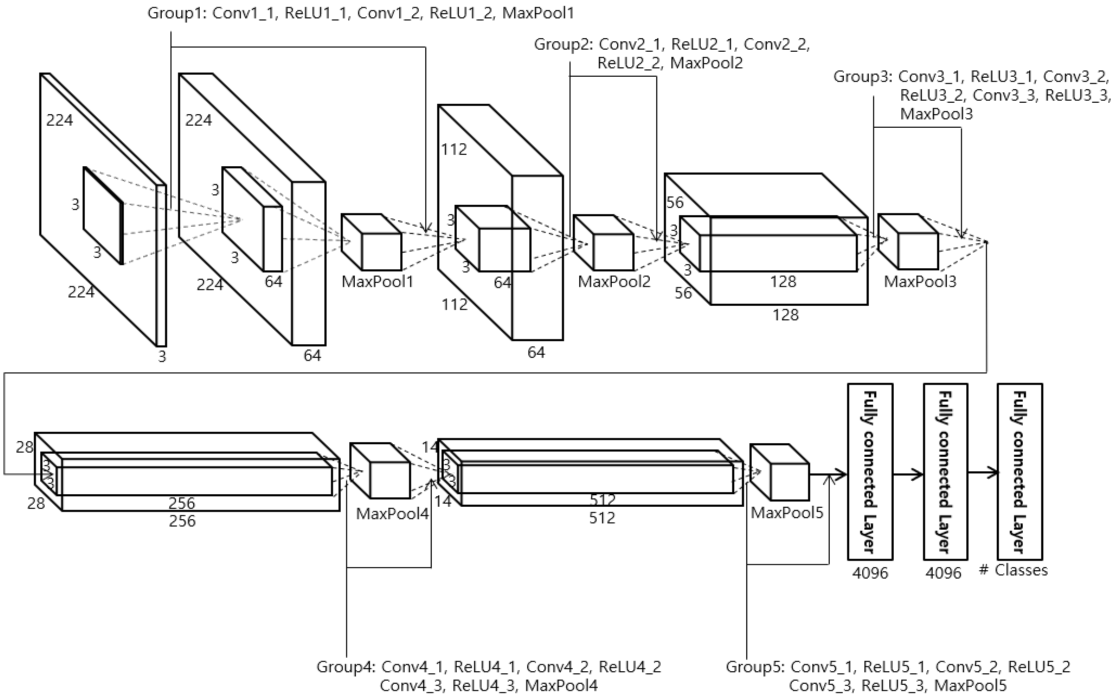

Group1: Conv1_1, ReLU1_1, Conv1_2, ReLU1_2, MaxPool1
Group2: Conv2_1, ReLU2_1, Conv2_2, ReLU2_2, MaxPool2
Group3: Conv3_1, ReLU3_1, Conv3_2, ReLU3_2, Conv3_3, ReLU3_3, MaxPool3
224
224
3
3
3
224
224
3
3
224
64
64
112
3
3
112
64
64
MaxPool1
MaxPool2
56
3
3
56
128
128
MaxPool3
28
28
256
256
MaxPool4
14
14
512
512
MaxPool5
Fully connected Layer
Fully connected Layer
Fully connected Layer
4096
4096
# Classes
Group4: Conv4_1, ReLU4_1, Conv4_2, ReLU4_2 Conv4_3, ReLU4_3, MaxPool4
Group5: Conv5_1, ReLU5_1, Conv5_2, ReLU5_2 Conv5_3, ReLU5_3, MaxPool5

Figure 6. The structure of VGG Face-16 [3]. Conv, ReLU, and MaxPool represent convolutional layer, rectified linear unit layer, and max pooling layer, respectively.

4.4. CNN for Human Recognition Using Body

The body region obtained in Section 4.2 is used as input for the ResNet-50 CNN to perform human recognition using body data. In this research we used the ResNet-50 CNN model provided in [69]. One of the ResNet-50 model's main features is the shortcut structure for residual learning shown in Figure 7 [70]. ResNet has many convolutional layers, so the feature map size becomes smaller the farther back one goes, and the vanishing or exploding gradient problem occurs as the feature map's feature values become smaller. Therefore, the shortcut structure shown in Figure 7 is used. Also, ResNet forms a bottleneck structure. The reason for this is that using 1 × 1, 3 × 3, and 1 × 1 convolutions rather than two 3 × 3 convolutions can reduce the computation time [70]. Batch normalization is performed before activation function and after each convolution [70,71]. In this study, the pre-trained ResNet-50 was fine-tuned with the training data. This ResNet-50 structure is shown in Figure 8 and Table 3.

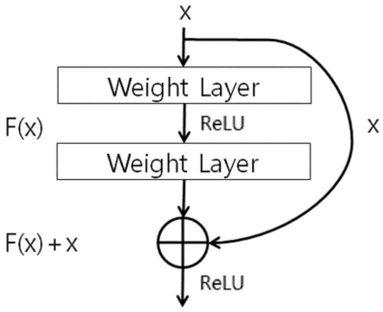


Figure 7. Shortcut for residual learning in ResNet. ReLU means rectified linear unit layer.

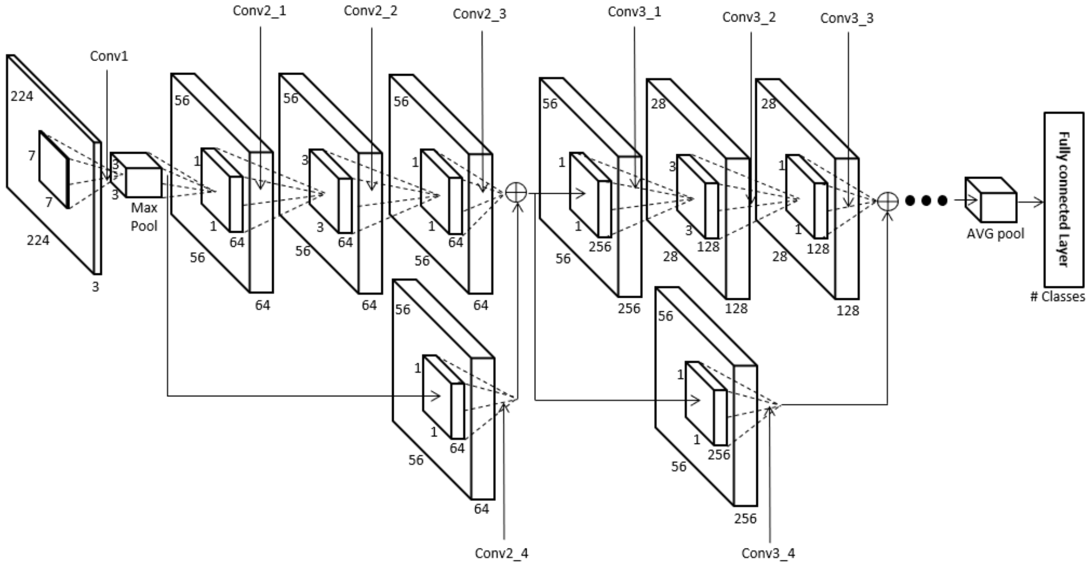


Figure 8. The structure of ResNet-50 [70]. Conv, MaxPool, and AVG pool represent convolutional layer, max pooling layer, and average pooling layer, respectively.

Table 3. Output size, numbers and sizes of filters, number of strides, and amount of padding in our deep residual CNN structure (3* indicates that 3 pixels are included as padding in left, right, up, and down positions of input image of 224 × 224 × 3, whereas 1* indicates that 1 pixel is included as padding in left, right, up, and down positions of feature map) (2/1** indicates 2 at the 1st iteration and 1 at the 2nd iteration) (For the shortcuts in Conv2_4, 3_4, 4_4, and 5_4, the filter of 1 × 1 is used only for the 1st iteration whereas identity mapping is used for the other iterations).

Layer Type		Size of Feature Map	Number of Filters	Size of Filters	Number of Strides	Amount of Padding	Number of Iterations
Image input layer		224 (height) × 224 (width) × 3 (channel)					
Conv1		112 × 112 × 64	64	7 × 7	2	3*	1
Max pool		56 × 56 × 64	1	3 × 3	2	0	1
Conv2	Conv2_1	56 × 56 × 64	64	1 × 1	1	0	3
	Conv2_2	56 × 56 × 64	64	3 × 3	1	1*	
	Conv2_3	56 × 56 × 256	256	1 × 1	1	0	
	Conv2_4 (Shortcut)	56 × 56 × 256	256	1 × 1	1	0	
Conv3	Conv3_1	28 × 28 × 128	128	1 × 1	2/1**	0	4
	Conv3_2 (Bottleneck)	28 × 28 × 128	128	3 × 3	1	1*	
	Conv3_3	28 × 28 × 512	512	1 × 1	1	0	
	Conv3_4 (Shortcut)	28 × 28 × 512	512	1 × 1	2	0	
Conv4	Conv4_1	14 × 14 × 256	256	1 × 1	2/1**	0	6
	Conv4_2 (Bottleneck)	14 × 14 × 256	256	3 × 3	1	1*	
	Conv4_3	14 × 14 × 1024	1024	1 × 1	1	0	
	Conv4_4 (Shortcut)	14 × 14 × 1024	1024	1 × 1	2	0	
Conv5	Conv5_1	7 × 7 × 512	512	1 × 1	2/1**	0	3
	Conv5_2 (Bottleneck)	7 × 7 × 512	512	3 × 3	1	1*	
	Conv5_3	7 × 7 × 2048	2048	1 × 1	1	0	
	Conv5_4 (Shortcut)	7 × 7 × 2048	2048	1 × 1	2	0	
AVG pool		1 × 1 × 2048	1	7 × 7	1	0	1
FC layer		2					1
Softmax		2					1

4.5. Training of CNN Model by Stochastic Gradient Descent Method

The stochastic gradient descent (SGD) method was used to train the VGG Face-16 and ResNet-50 used in this paper. SGD is a type of gradient descent method, and it is expressed as [72]:

$$W_{n+1} = W_n - \gamma \nabla \mathrm{F}(W_n), \tag{3}$$

where W represents the parameters of the CNN which must be found via training. It consists of the product of the movement distance γ from the activation function F(x), which takes the value of the previous parameters as input. Depending on whether the initial starting point is a negative number or positive number, $\gamma\nabla\mathrm{F}(x)$ amount of movement is made in the opposite direction. Unlike the gradient descent (GD) method, which uses all the training data to find the optimal parameters, in the SGD method training is performed in mini-batch units (Z of Equation (4)) randomly selected from the overall training data [72]:

$$W_{n+1} = W_n - \gamma \nabla \mathrm{F}(Z_n, W_n). \tag{4}$$

The codes of SGD method for VGG Face-16 and ResNet-50 are provided from [61,69], respectively. The detail parameters for SGD method used in our experiments are explained in Section 5.2.

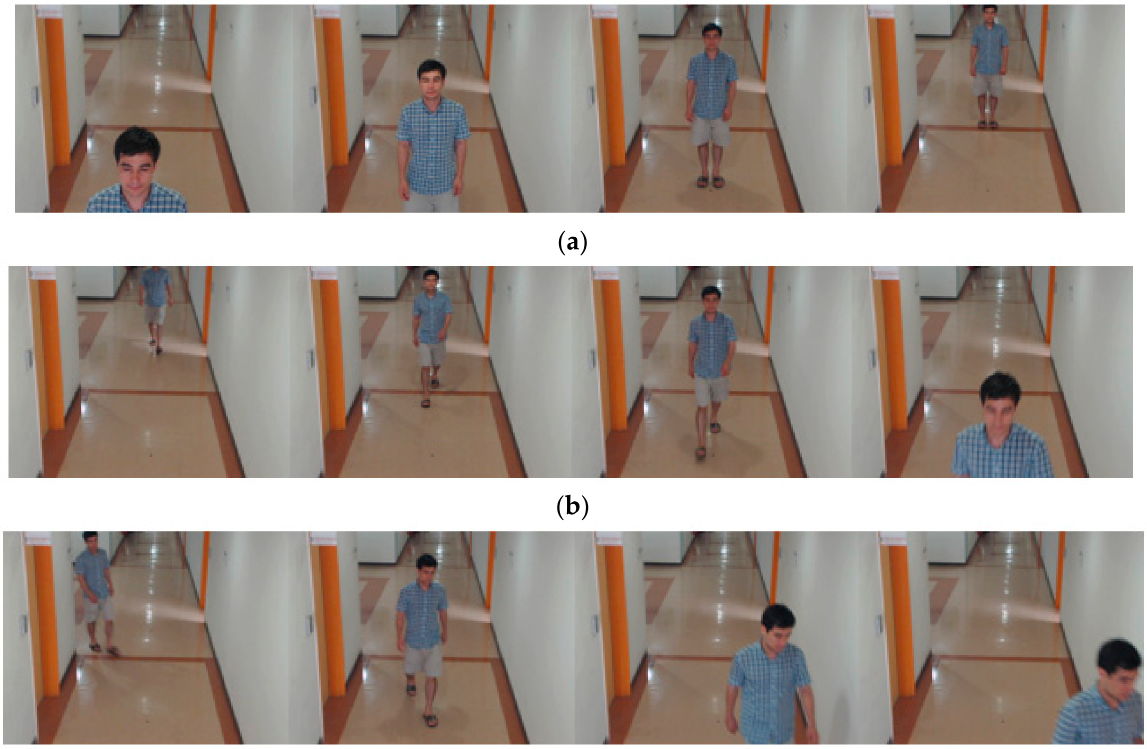
4.6. Calculation of Distance and Score-Level Fusion

In the next step, the 4096 features behind the 2nd fully connected layer in Table 2 are used as features for face recognition, and the 2048 features behind the AVG pool in Table 3 are used as features for human recognition using body. After this, we find each of the Euclidean distances from the features previously extracted from the enrolled images. The two Euclidean distances are normalized through min–max scaling, and score-level fusion is performed to find the final matching score. Here, the min and max values for min–max scaling are found in the training data. For score-level fusion, the weighted sum and weighted product rules are used. For score level fusion, two scores from face and human recognition using body are normalized via min-max scaling, and optimal weights for score level fusion were found from the training data. Based on the fused score, recognition is performed. In detail, in case of verification (1:1 matching), if the fused score is less than the predetermined threshold, the input image is accepted as genuine matching. If not, it is rejected as imposter matching. Here, the genuine matching means the case that input and enrolled images are from a same class whereas the imposter matching represents the case that input and enrolled images are from a different class. The optimal threshold was experimentally determined with training data so as to obtain the minimum equal error rate (EER) of recognition. There are two types of error rates such as false acceptance rate (FAR) and false rejection rate (FRR). These two error rates have the trade-off relationship. That is, the larger the FAR, the smaller the FRR. The EER is the error rate when FAR is same to FRR. In case of identification (1:n matching), one enrolled image (among n images) which shows the smallest fused score with the input is determined as that of same class to the input image.

5. Experimental Results and Analysis

5.1. Experimental Data and Environment

In this study, DFB-DB1 was created for the experiments using images of 22 people obtained by two types of cameras to assess the performance of the proposed method in a variety of camera environments. The first camera was a Logitech BCC 950 [58], and the camera specifications include a camera viewing angle of 78°, a maximum resolution of full high-definition (HD) 1080 p, and auto-focusing at 30 frames per second (fps). The second camera was a Logitech C920 [59], and its specifications include a maximum resolution of full HD 1080p, a viewing angle of 78° at 30 fps, and auto focusing. Images were taken in an indoor environment with indoor lights on, and each camera was installed at a height of 2 m 40 cm. Before collecting DFB-DB1, we gave the sufficient explanations of our experiments to acquire DFB-DB1 to all the participants. In addition, we obtained the informed and signed consent forms from all the participants before collecting DFB-DB1, and all the participants also agreed to show their faces and bodies (without any pre-processing) in our paper. The database was divided into two categories according to the camera. In the first database, the images were captured by the Logitech BCC 950 based on the scenarios of one, two, and three people, including the images of two cases where the target body was still and when it was moving. The still images were captured in four positions, and the moving images were divided into two cases (straight-line movement and corner movement) and captured. We requested all the participants to move naturally without noticing the situation of collecting our DFB-DB1, and did our best for collecting DFB-DB1 in the real-world scenario. Examples of still images and movement images are shown in Figure 9. The second database is composed of the images obtained by the Logitech C920, and the angle of camera was similar to that for capturing the first database. In the second database, the images were captured based on the scenario of 1 people and the case where the target body was moving (straight-line movement) by three times, as shown in Figure 10.


(**a**)

(**b**)

(**c**)

Figure 9. Example images from DFB-DB1 taken by the Logitech BCC 950 camera. (**a**) One-person still image. (**b**) One-person straight-line movement image. (**c**) One-person corner movement image.

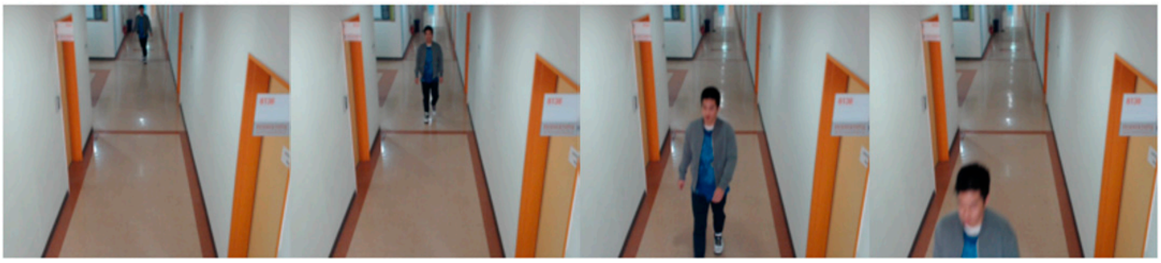


Figure 10. Example images from DFB-DB1 taken by the Logitech C920 camera.

Table 4 contains a description of DFB-DB1. This study executed a two-fold cross validation scheme, so DFB-DB1 was divided into sub-databases 1 and 2. In the first cross validation, sub-database 1 was used for training and sub-database 2 was used for testing. In the 2nd fold cross validation, sub-database

2 was used for training, and sub-database 1 was used for testing. Sub-databases 1 and 2 were made to contain images of different people. Also, DFB-DB1 and the VGG Face-16 and ResNet-50 models which were trained in this study were made public for other researchers in [53] so that impartial comparison experiments could be performed.

The ChokePoint database is a real-world surveillance video database which was designed for person identification and verification experiments and is provided by National ICT Australia Ltd. (NICTA) as an open database [60]. It consists of Portals 1 and 2. Portal 1 contains images of 25 people (19 males and 6 females), and Portal 2 contains images of 29 people (23 males and six females). Portals 1 and 2 were captured during a one-month time span. The images for each location were captured with three cameras, and at a total of six locations. In this study, the location P2L was selected from among the six locations as it is similar to the location in the DFB-DB images. As previously mentioned, the P2L database contains images of a total of 29 people. In this study 28 people were selected for two-fold cross validation. Fourteen classes were set for each of the sub-databases 1 and 2. Examples from the ChokePoint database are shown in Figure 11, and descriptions of the ChokePoint database are provided in Table 4.

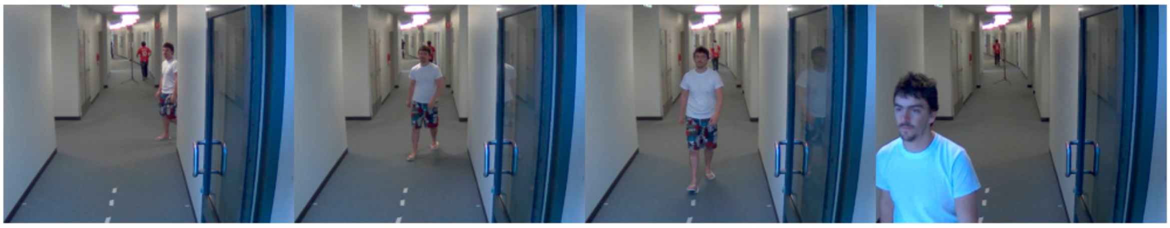


Figure 11. ChokePoint dataset image example.

Table 4. Descriptions of DFB-DB1 and ChokePoint dataset.

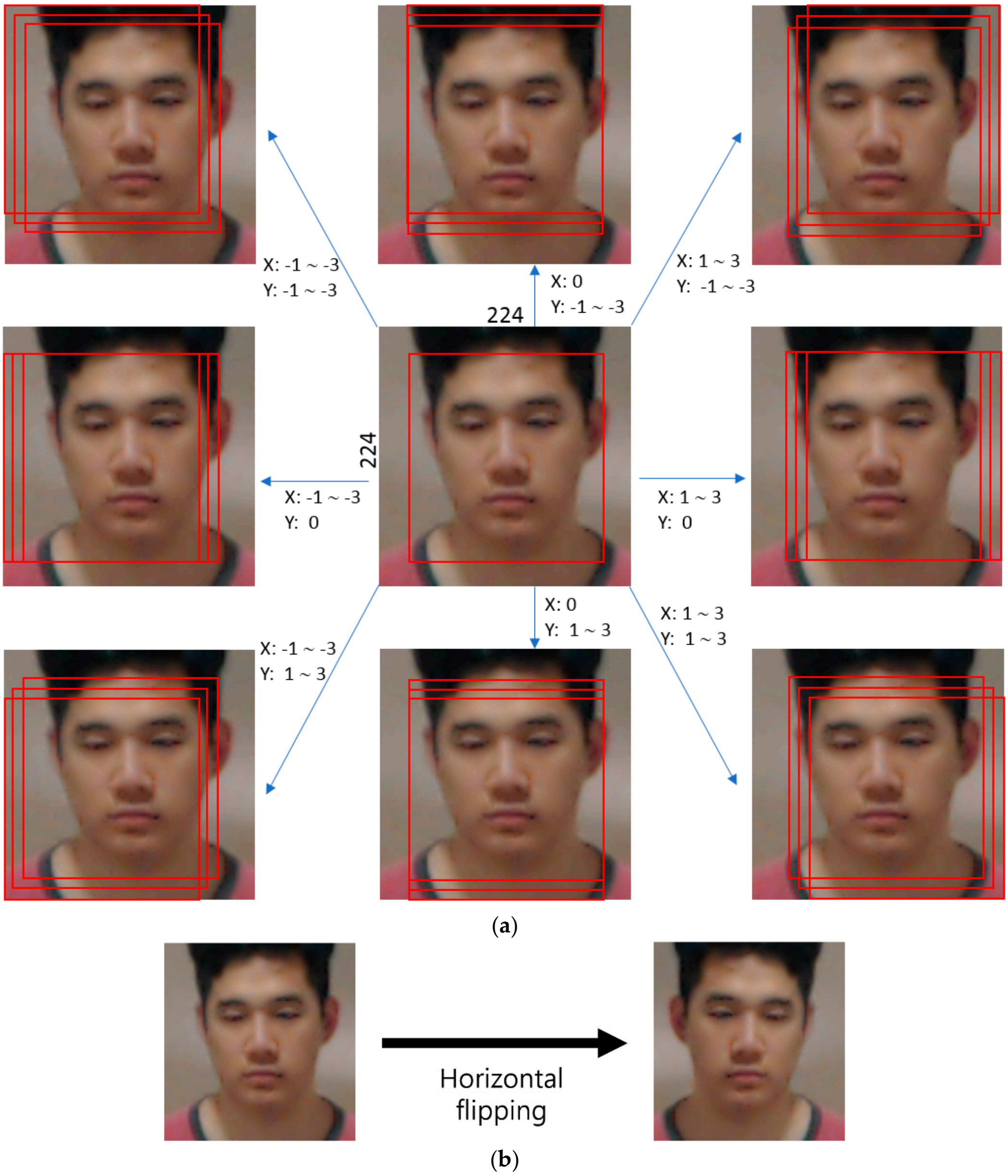
		Face		Body	
		Sub-Dataset 1	**Sub-Dataset 1**	**Sub-Dataset 1**	**Sub-Dataset 1**
DFB-DB1	Number of people	11	11	11	11
	Number of images	564	767	564	767
	Number of augmented images (for training)	278,300	324,038	278,300	324,038
ChokePoint dataset	Number of people	14	14	14	14
	Number of images	7565	7296	7565	7296
	Number of augmented images (for training)	378,250	364,800	378,250	364,800

In this study, the training and tests were performed in a desktop environment that included an Intel Core i7-6700 CPU @ 3.4 GHz (four cores) with 16 GB of RAM, and NVIDIA GeForce GTX 1070 with a graphics memory of 8 GB [73] (CUDA 8.0). The Windows Caffe framework (version 1) [74], Microsoft Visual Studio 2013 [75], and OpenCV library (ver. 2.4.10) [56] were used to implement the algorithm.

5.2. Training of CNN Model

To resolve the problem of the CNN not receiving adequate training due to insufficient training data, training in this study was performed using data that was increased through the augmentation of the training data using the method described below. As shown in Table 4, data augmentation was performed only on the training data, and only unaugmented original data was used for the testing data.

In DFB-DB1, the number of images for each class (person) is different, so when augmentation was performed, classes with over 100 images underwent a process of 3-pixel left/right/top/bottom image translation and cropping as well as horizontal flipping (mirroring) (refer to Figure 12), while classes with less than 100 images underwent a process of 5-pixel left/right/top/bottom image translation and cropping as well as horizontal flipping. Sub-databases 1 and 2 from Table 4 were combined to obtain around 600,000 augmented images. In the ChokePoint dataset, unlike DFB-DB1, there were many images for each class, so image translation and cropping was performed at 2-pixel increments in the upper-left direction and 2-pixel increments in the lower-right direction to increase the number of images by a factor of 25. In addition, a horizontal flipping process was performed to increase the number of images by a factor of 50. Sub-databases 1 and 2 from Table 4 were combined to obtain around 740,000 augmented images. This data augmentation method has been used many times in previous studies [76].

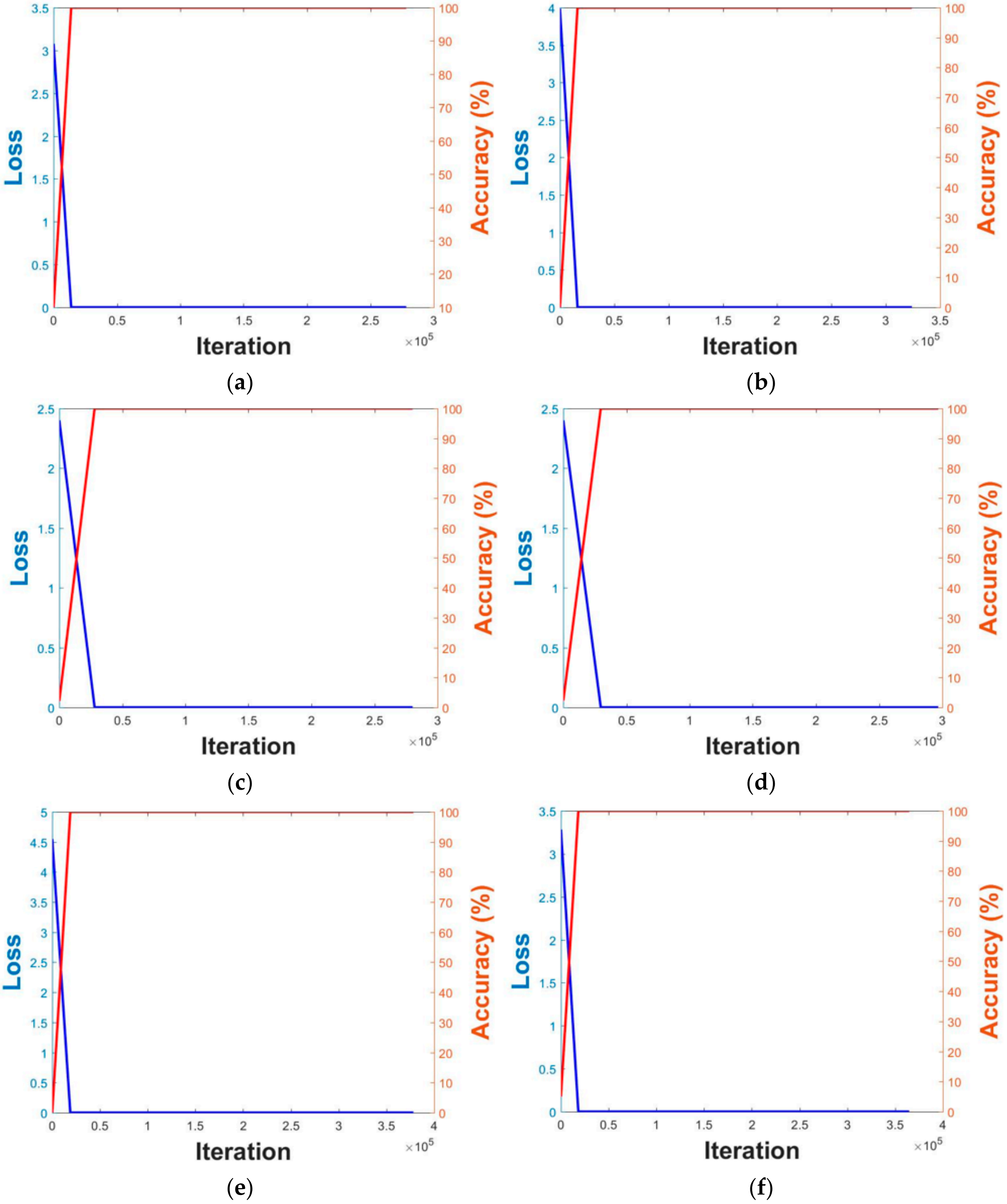


(**a**)

(**b**)

Figure 12. Data augmentation method by (**a**) image translation and cropping, and (**b**) horizontal flipping.

Using the augmented data, fine-tuning was performed on pre-trained VGG Face-16 and ResNet-50 models using the SGD method. As explained in Section 4.5, unlike the GD method, in the SGD method, the number of training sets divided by mini-batch size is defined as an iteration, and one epoch is set when training is performed for all the iterations. In this study, the momentum, weight decay, and learning rate during training were set at 0.9, 5×10^{-4}, and 1×10^{-5}, respectively, and the batch size was 20. Training with DFB-DB1 was performed for 20 epochs, and training with the ChokePoint database was performed for 15 epochs. Because the number of images in the ChokePoint database is larger than that in DFB-DB1 as shown in Table 4, CNN training with the ChokePoint database was performed by the smaller number of epochs than that in DFB-DB1 considering the limitation of graphic processing unit (GPU) memory. Figure 13 shows the training loss and accuracy during the 1st and 2nd validations using DFB-DB1 and the ChokePoint databases. The x axis shows the number of iterations, while the left-side of the y axis shows the loss value and the right-side of the y axis shows the training accuracy. As seen in Figure 13, the training loss was close to 0%, and the training accuracy was close to 100% in all cases. This indicates that the VGG Face-16 and ResNet-50 models used in this study were sufficiently trained. Experimental results showed that it took two or three days for training one model in each fold.

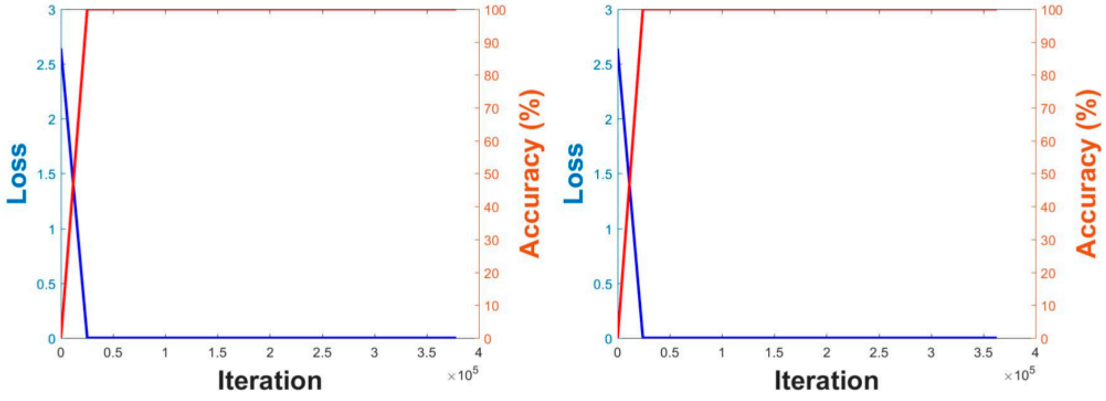


(a) (b)

(c) (d)

(e) (f)

Figure 13. *Cont.*



(**g**) (**h**)

Figure 13. Graphs of training loss and accuracy on DFB-DB1 (**a**–**d**) and ChokePoint datasets (**e**–**h**). VGG Face-16 in case of (**a**,**e**) the 1st validation, (**b**,**f**) the 2nd validation. ResNet-50 in case of (**c**,**g**) the 1st validation, (**d**,**h**) the 2nd validation. Red and blue lines show the training accuracy and loss, respectively.

5.3. Testing of Proposed Method

5.3.1. Comparisons of Accuracy Achieved by VGG Face-16 and ResNet-50 for Face or Body Recognition

The first experiment measured the accuracy of the VGG Face-16 face recognition and the ResNet-50 body recognition. An equal error rate (EER) was found from the authentic and imposter matching distribution, which was based on the Euclidean distance between the enrolled and input images calculated based on the 4096 features of VGG Face-16. Also, an EER was found from the authentic and imposter matching distribution, which was based on the Euclidean distance between the enrolled and input images calculated based on the 2048 features of ResNet-50. Authentic matching occurs when the enrolled and input images are images of the same class, and imposter matching occurs when they are images of different classes. Also, an error in which an authentic match is incorrectly rejected as an imposter match is called a false rejection error (FRR). Conversely, an error in which an imposter match is incorrectly accepted as an authentic match is called a false acceptance error (FAR). FRR and FAR have a trade-off relationship with each other, and the point at which the FAR and FRR rates become the same is called the equal error rate (EER). As mentioned earlier, experiments were performed with two-fold cross validation using the mean error obtained from testing two times.

First, to compare the recognition accuracy of each CNN model in the face and body regions, the EER of VGG Face-16 and ResNet-50 in testing after training was measured for facial recognition and body recognition, as shown in Tables 5 and 6, respectively. As seen in the tables, VGG Face-16 made fewer errors in face recognition, and ResNet-50 made fewer errors in body recognition. This suggests that ResNet-50, which has more layers and uses detailed residual information, showed better performance in the body region because the body region is larger than the face region and detailed texture, color, and shape data must be extracted from the clothes and body. Conversely, VGG Face-16 showed better performance than ResNet-50 in the face region because the face region is smaller than the body region, and normally mid- or low-frequency information is used in recognition rather than high-frequency information.

Table 5. Comparisons of EERs by VGG Face-16 and ResNet-50 for face recognition (unit: %).

	VGG Face-16	**ResNet-50 [40,41]**
1st fold	2.03	9.11
2nd fold	2.49	17.7
Average	2.26	13.405

Table 6. Comparisons of EERs by VGG Net-19 and ResNet-50 for body recognition (unit: %).

	VGG Net-19 [42,43]	ResNet-50
1st fold	27.52	8.82
2nd fold	16.21	7.88
Average	21.865	8.35

5.3.2. Comparisons of Accuracy Achieved by Single Modality-Based Method and Score-Level Fusions

In the next experiment, the accuracy of single modality-based recognition, which uses face and human recognition using body individually, was compared with the accuracy of the score-level fusion used in this study. For score-level fusion, the weighted sum and weighted product methods described in Section 4.6 were compared. As seen in Tables 7 and 8, the weighted sum method achieved higher accuracy than the weighted product method in both databases, and it achieved higher accuracy than single modality-based recognition of the face and body without score-level fusion. That is because the two dimensional classifier based the two scores of face and human recognition using body is used for classification in case of score-level fusion whereas one dimensional classifier is used for single modality-based recognition.

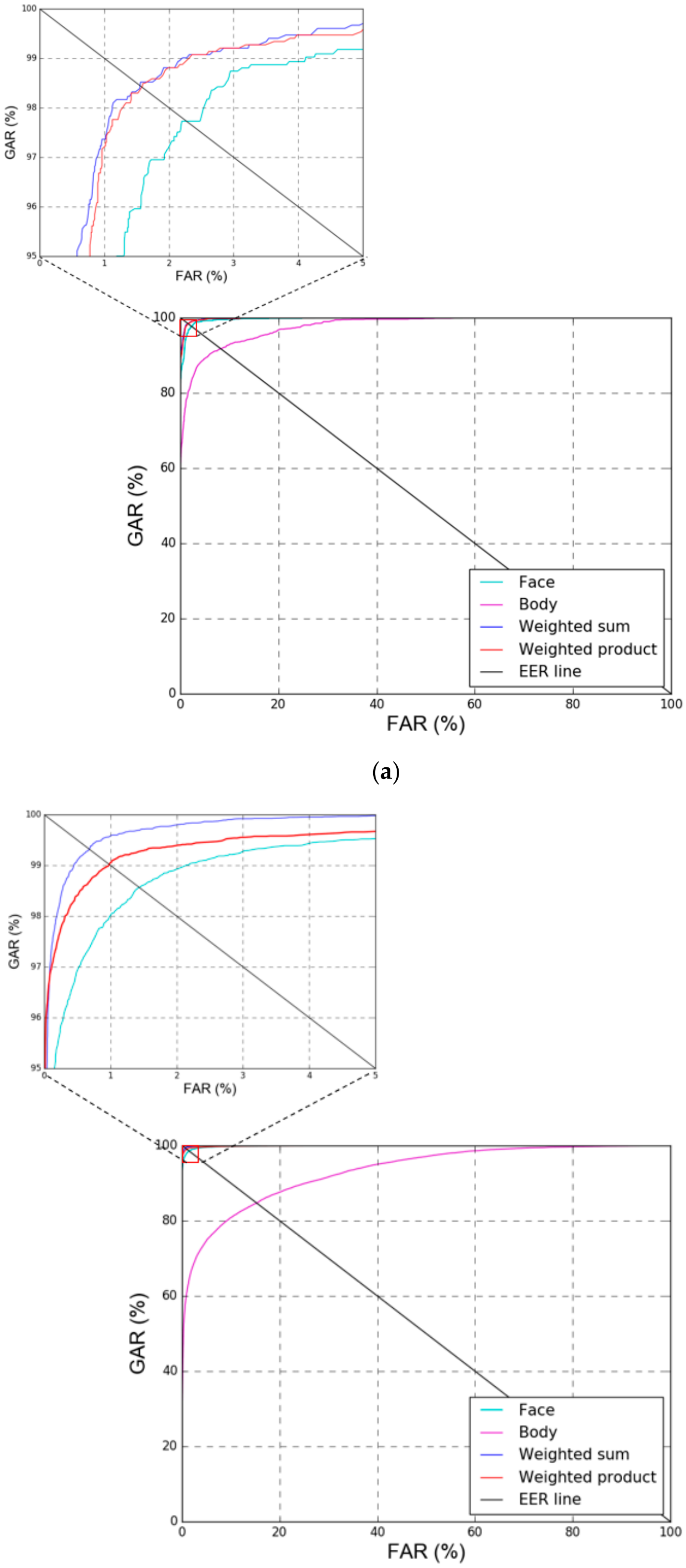
Figure 14 shows the receiver operating characteristic (ROC) curves [77] of the results of Tables 7 and 8. Here, the genuine acceptance rate (GAR) is defined as 100-FRR (%). As previously mentioned, the experiments in this study were performed with two-fold cross validation, and the average graph of the ROC curve obtained from testing two times is shown. In Figure 14, it can be seen that the weighted sum method showed higher accuracy than the weighted product method in both databases, and it showed higher accuracy than single modality-based recognition of the face and body without score-level fusion.

Table 7. Comparisons of EERs by face and human recognition using body (unit: %).

Modality	DFB-DB1			ChokePoint Dataset		
	1st Fold	2nd Fold	Average	1st fold	2nd Fold	Average
Face	2.03	2.49	2.26	1.49	1.38	1.435
Body	8.82	7.88	8.35	18.44	10.67	14.56

Table 8. Comparisons of EER by score-level fusion (unit: %).

Method	DFB-DB1			ChokePoint Dataset		
	1st Fold	2nd Fold	Average	1st Fold	2nd Fold	Average
Weighted Sum	0.9	2.13	1.52	0.37	0.79	0.58
Weighted Product	0.92	2.23	1.58	1.12	0.88	1



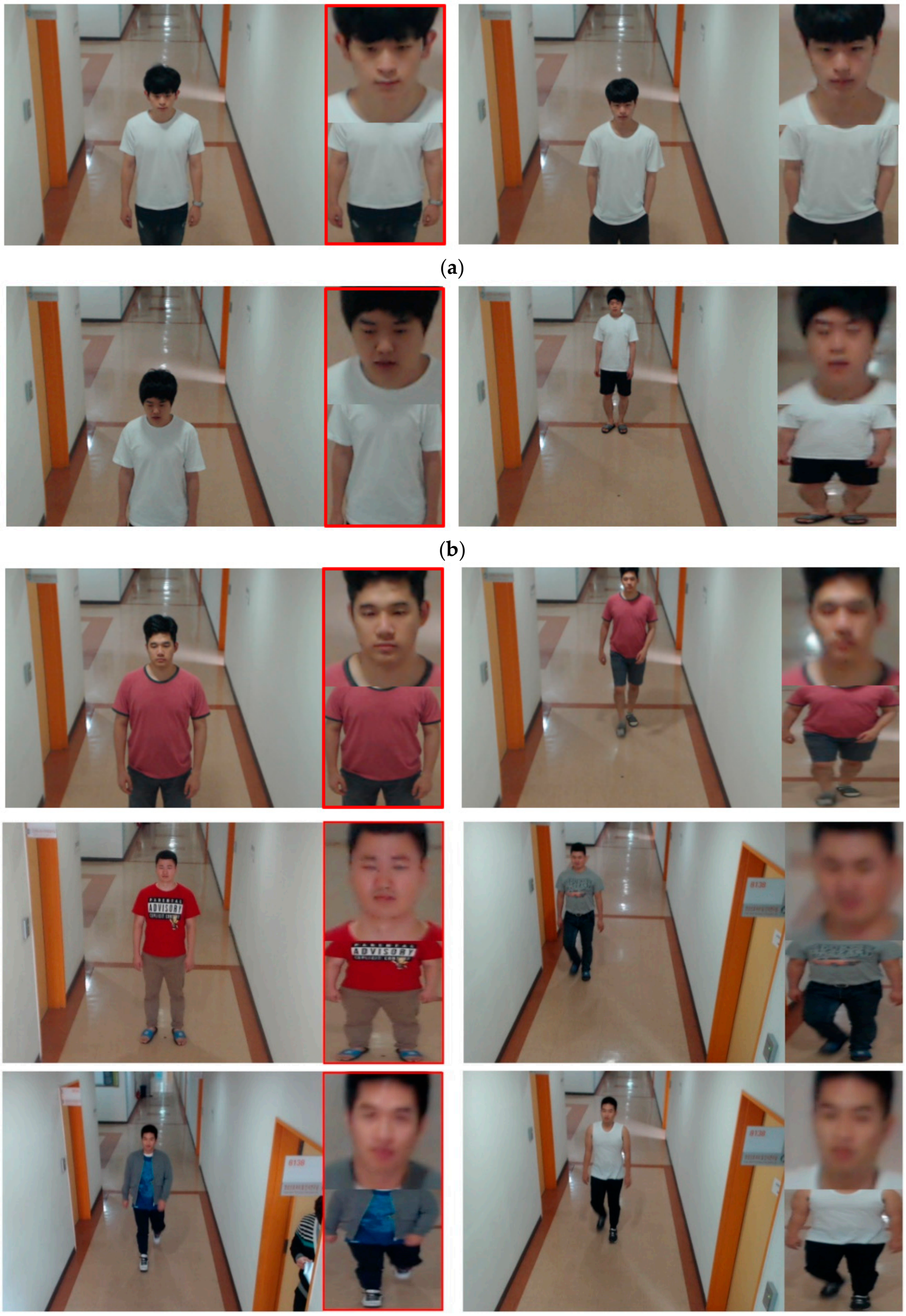
(**a**)

(**b**)

Figure 14. ROC curves by single modality-based method and score-level fusion. With (**a**) DFB-DB1, (**b**) ChokePoint dataset. GAR, FAR, and EER mean genuine acceptance rate, false acceptance rate, and equal error rate, respectively.

5.3.3. Cases of Correct Recognition, False Acceptance (FA), and False Rejection (FR)

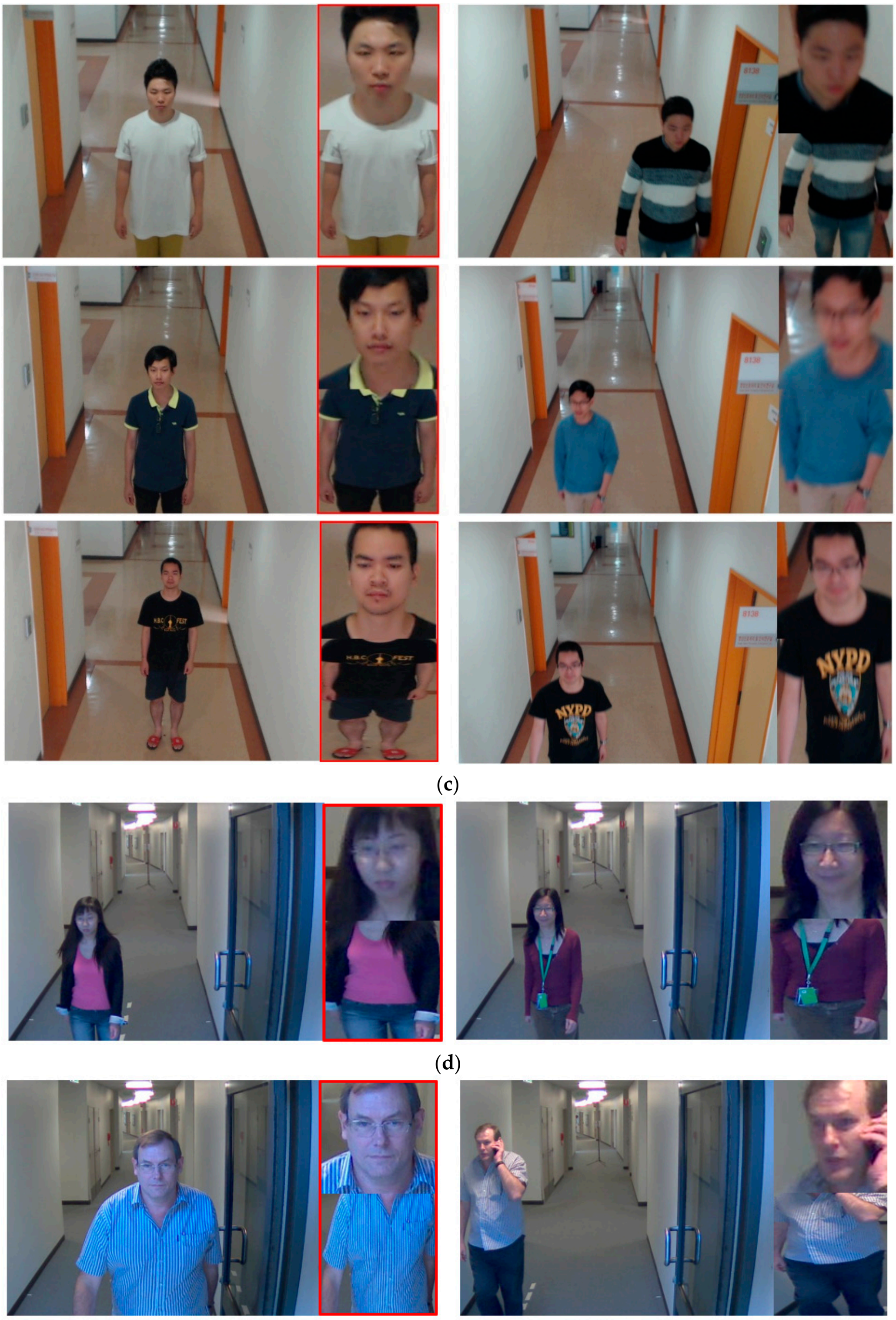
In this section, we present cases of correct recognition, false acceptance, and false rejection as shown in Figure 15. The image in the red box on the left side of Figure 15 is the enrolled image, and the image on the right side is the recognition image.



(**a**)

(**b**)

Figure 15. *Cont.*

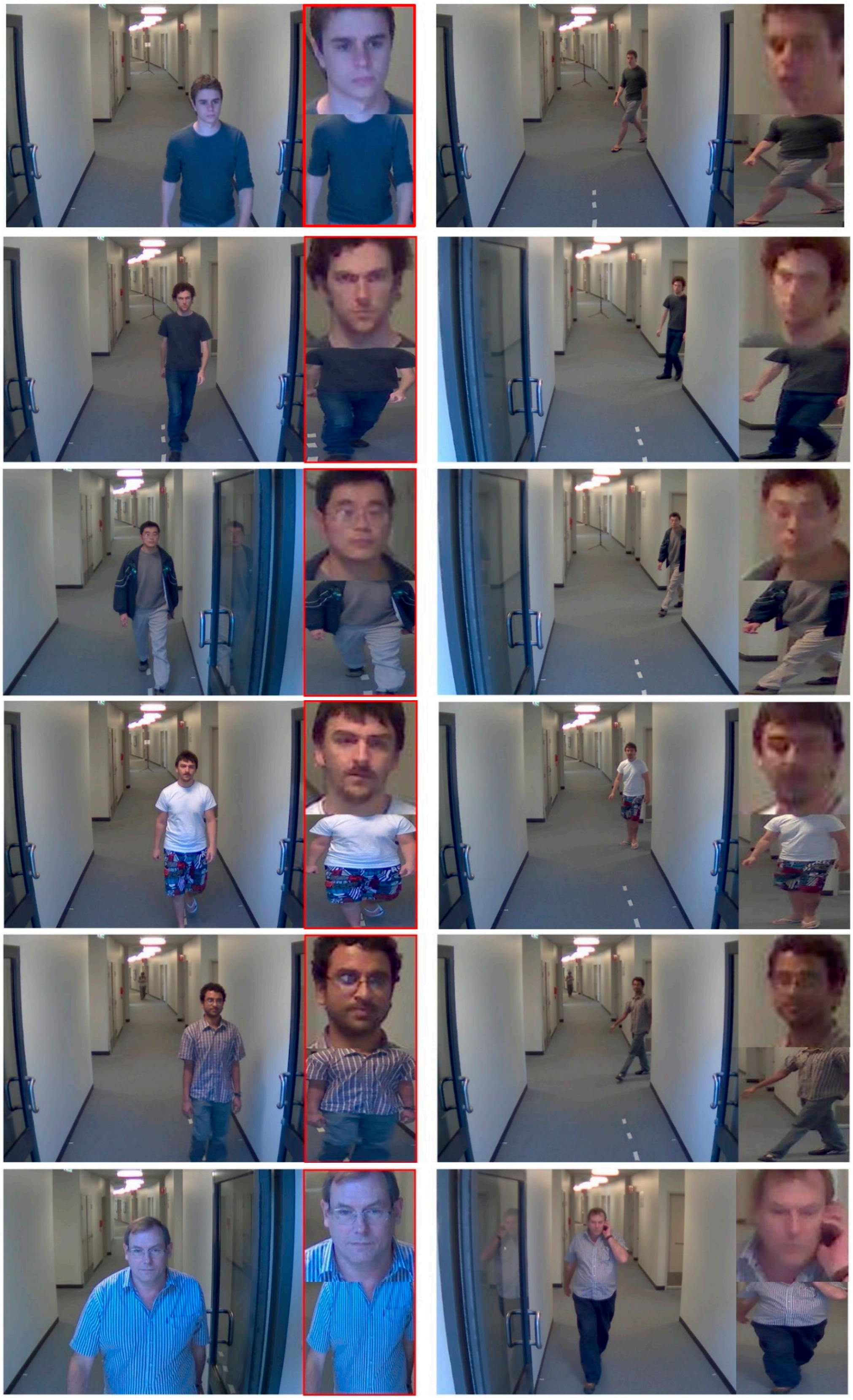


(**c**)

(**d**)

(**e**)

Figure 15. *Cont.*



(**f**)

Figure 15. Cases of false acceptance (FA), false rejection (FR), and correct recognition. (**a**–**c**) cases from Dongguk face and body database (DFB-DB1), (**d**–**f**) cases from ChokePoint dataset. (**a**,**d**) FA cases. (**b**,**e**) FR cases. (**c**,**f**) cases of correct recognition.

As seen in Figure 15a,d, FA occurred when the face and body shapes were similar even thought it was an imposter. Also, as shown in Figure 15b,e, FR occurred when the face was blurred, when a hand and mobile phone were partially included in the face region, when changes in the face pose occurred, and when there was a big difference in the body shape between the enrolled image and the recognition image (when legs were only included in the recognition image and changes in the body's pose had occurred). However, as Figure 15c demonstrates, even when there was face blurring, the difference in body shape and size between the enrolled image and the recognition image, correct recognition results were achieved by the method proposed in this study. As shown in the 2nd to 6th row images of Figure 15c, the same people even with different clothes were correctly recognized by our system. That is because the person's whole body image is not used as a single CNN input. Rather, the face region and the body region are separated, and each is used as a separate CNN input. Therefore, the difference of clothes can be compensated by face recognition. In particular, if we disregard the case shown in Figure 15f, where the recognition image is captured at a long distance at the moment the person is coming around a corner and the face image's resolution is very poor and there are large changes in body shape and pose, the correct recognition results were achieved through score-level fusion of the 2 deep CNN results that were used in this study.

5.3.4. Comparison of Recognition Accuracy by Proposed Method and Using One CNN Based on Full Body Image, and That with and without Data Augmentation

In the next experiment, a performance comparison was made between the method proposed in this study, in which face and body regions are separately processed by two CNNs and score-level fusion is performed, and a method which performs recognition based on one CNN that uses the face and body regions in a single input image. For experiments, VGG Face-16 and ResNet-50 models were fine-tuned with our experimental images. As seen in Table 9, the method proposed in this study achieved higher recognition accuracy. It was possible to use the method to recognize more detailed texture, color, and shape data in each region by separating the face and body regions and using them as input in separate CNNs.

Table 9. Comparisons of EERs by proposed method and using one CNN based on full body image (unit: %).

	Using One CNN Based on Full Body Image (VGG Face-16) [42]	**Using One CNN Based on Full Body Image (ResNet-50)**	**Proposed Method**
1st fold	12.49	4.98	0.9
2nd fold	13.59	2.65	2.13
Average	13.04	3.815	1.52

As the next experiment, we compared the accuracy of the models with and without data augmentations. For fair comparison, same procedure of two-fold cross validation was adopted for both methods with and without data augmentations. As shown in Table 10, the EER of recognition with data augmentation is much lower than that without augmentation. The reason why the EER becomes higher without data augmentation is that the number of data is insufficient for training our deep CNN.

Table 10. Comparisons of EERs with and without data augmentation (unit: %).

	With Augmentation			**Without Augmentation**		
	Face	**Body**	**Combined**	**Face**	**Body**	**Combined**
1st fold	2.03	8.82	0.9	13.56	50.32	13.53
2nd fold	2.49	7.88	2.13	12.6	17.24	10.3
Average	2.26	8.35	1.52	13.08	33.78	11.92

As the next experiment, we included the analysis of the influence of focus assessment on the next steps of proposed method. For that, we performed the additional experiments to measure the recognition accuracies with and without focus assessment. For fair comparison, same procedure of two-fold cross validation was adopted for both methods with and without focus assessment. As shown in Table 11, our method with focus assessment shows much lower errors of recognition compared to that without focus assessment. Without focus assessment, severely blurred images are attempted to be recognized, which increases the errors of recognition.

Table 11. Comparisons of EERs with and without focus assessment (unit: %).

	With Focus Assessment			Without Focus Assessment		
	Face	Body	Combined	Face	Body	Combined
1st fold	2.03	8.82	0.9	47.19	32.58	29.67
2nd fold	2.49	7.88	2.13	47.09	28.94	26.42
Average	2.26	8.35	1.52	47.14	30.76	28.05

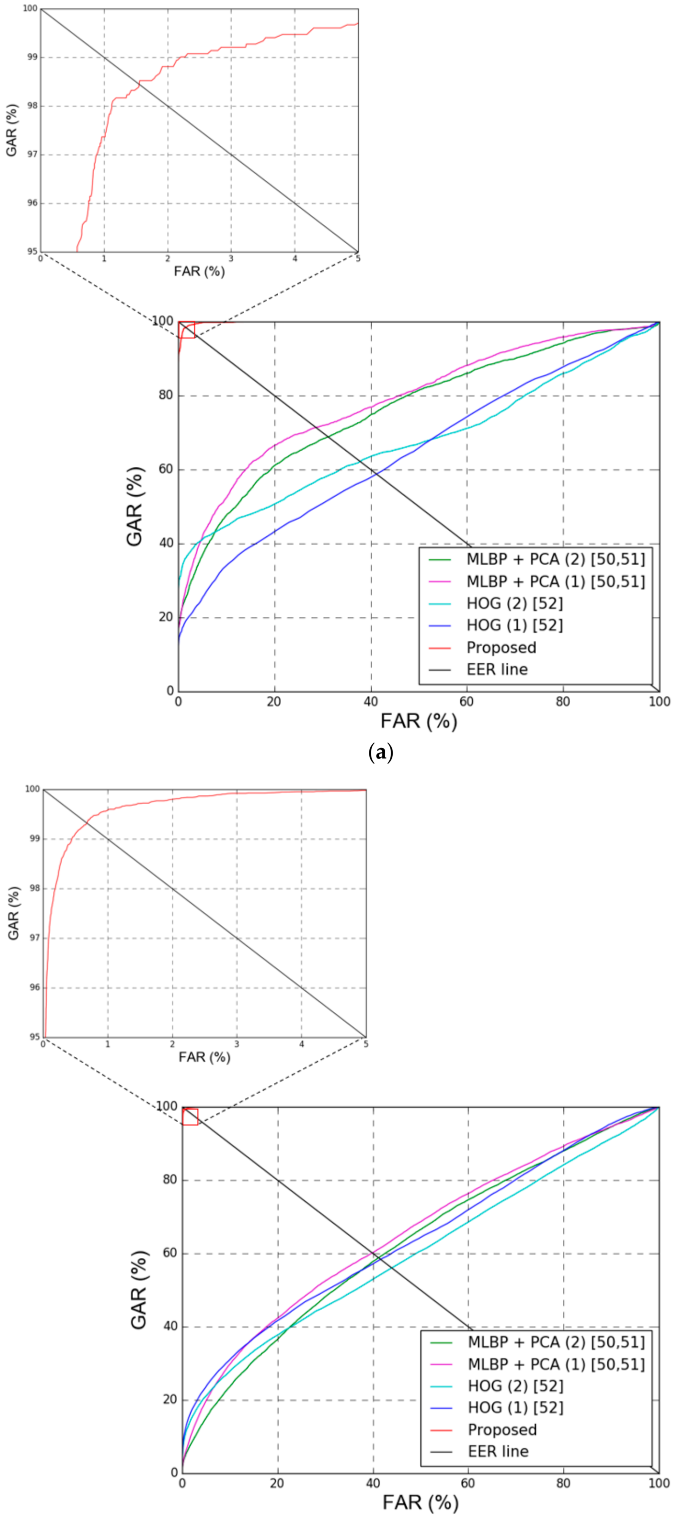
5.3.5. Comparisons of Accuracies by Proposed and Previous Methods

The next experiment compared the recognition accuracy of the proposed method and that of previous methods based on HOG [52] and multi-level local binary pattern (MLBP) + PCA [50,51]. When the accuracy of previous methods was assessed, the methods were divided into two types according to the way of determining enrolled images, and the experiments were performed.

The first type determines enrolled images by assuming that the image with the smallest mean value for the image pixel difference with different images in the same class is the geometric center of the feature space. The second type determines enrolled images by assuming that the image with the smallest mean value for the feature difference with different images in the same class is the geometric center of the feature space. For fair comparison, same procedure of two-fold cross validation was adopted for all the experiments. As shown in Table 12 and Figure 16, the other methods all have lower recognition accuracy than the proposed method.

Table 12. Comparison of EERs by proposed and previous methods (unit: %).

Method		DFB-DB1			ChokePoint Dataset		
		1st Fold	2nd Fold	Average	1st Fold	2nd Fold	Average
HOG [52]	Geometric center by pixel difference	40.13	35.67	37.9	45.98	42.19	44.09
	Geometric center by feature difference	38.09	44.14	41.12	41.84	41.47	41.66
MLBP + PCA [50,51]	Geometric center by pixel difference	31.72	30.62	31.17	41.92	39.9	40.91
	Geometric center by feature difference	29.38	27.84	28.61	37.75	42.38	40.07
Proposed method		0.9	2.13	1.52	0.37	0.79	0.58

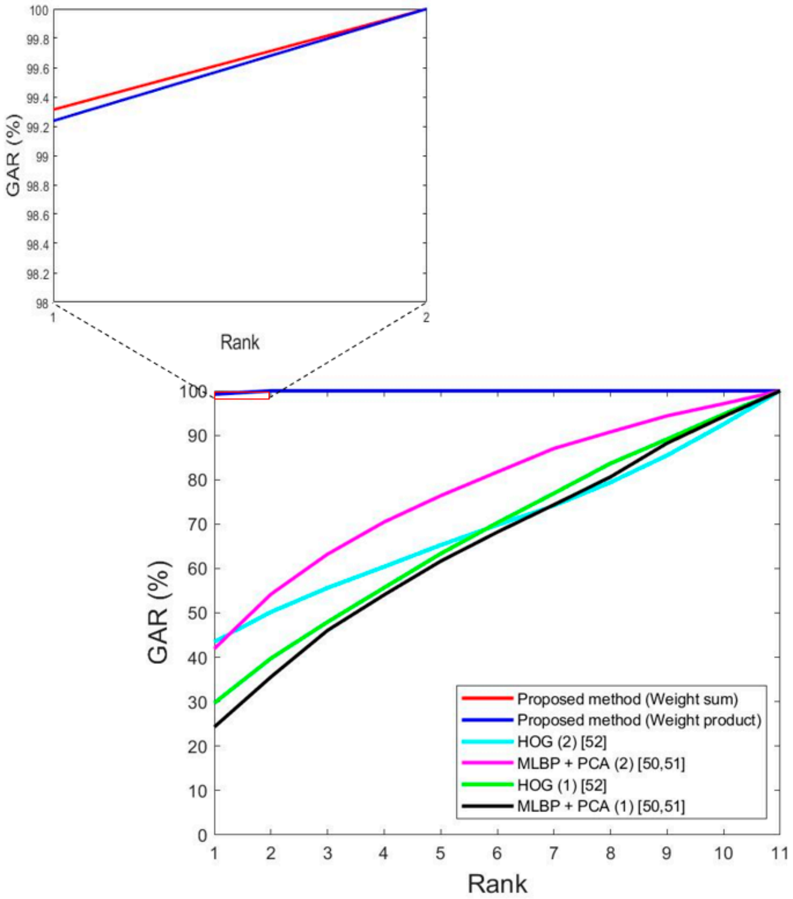


(**a**)

(**b**)

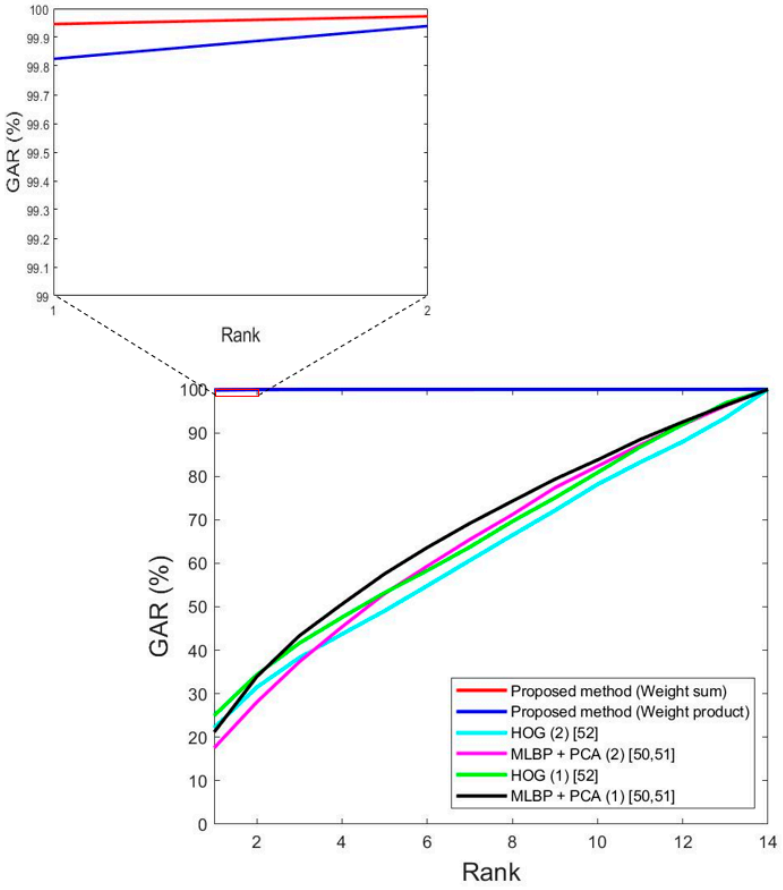
Figure 16. ROC curves by proposed and previous methods. With (**a**) Dongguk face and body database (DFB-DB1) and (**b**) ChokePoint datasets. In (**a**,**b**), multi-level local binary pattern (MLBP) + principal component analysis (PCA) (1) and MLBP + PCA (2) mean the methods of MLBP + PCA based on geometric center by feature difference and pixel difference of Table 12, respectively. In addition, histogram of oriented gradient (HOG) (1) and HOG (2) mean the methods of HOG based on geometric center by feature difference and pixel difference of Table 12, respectively. GAR, FAR, and EER mean genuine acceptance rate, false acceptance rate, and equal error rate, respectively.

The next experiment measured the cumulative match characteristic (CMC) curve to evaluate identification accuracy. Figure 17 shows the CMC curves. The horizontal axis shows the rank, and the vertical axis shows the accuracy (GAR) by rank. As shown in Table 4, 11 people's data are included in both sub-datasets 1 and 2 for DFB-DB1, and the maximum rank becomes 11 as shown in Figure 17a. In addition, as shown in Table 4, 14 people's data are included in both sub-datasets 1 and 2 for ChokePoint datasets, and the maximum rank becomes 14 as shown in Figure 17b. As an example, the meaning of a 90% GAR at rank 10 is that when the enrolled image with the smallest matching distance to the input image is selected, the case where the selected image is included in the 10 candidates based on matching distance rank is considered a genuine acceptance case, and the accuracy of this is 90%.



(**a**)

Figure 17. *Cont.*



(**b**)

Figure 17. CMC curves by proposed and previous methods. With (**a**) Dongguk face and body database (DFB-DB1) and (**b**) ChokePoint dataset. In (**a**,**b**), multi-level local binary pattern (MLBP) + principal component analysis (PCA) (1) and MLBP + PCA (2) mean the methods of MLBP + PCA based on geometric center by feature difference and pixel difference of Table 12, respectively. In addition, histogram of oriented gradient (HOG) (1) and HOG (2) mean the methods of HOG based on geometric center by feature difference and pixel difference of Table 12, respectively.

As shown in Table 4, there were 11 people in the DFB-DB1 database's testing sub-database and 14 people in the ChokePoint database's testing sub-database, so the horizontal axes in Figure 17a,b show 11 and 14. As seen in Figure 17, the accuracy of the proposed method was higher than that of previous methods in terms of the CMC curves.

5.3.6. Discussion

Gait recognition with continuous images can show better accuracy than our single-image based approach combining face and body recognition. However, in most previous researches for gait recognition with continuous images [78–85], the accurate region and boundary of human body including the legs should be segmented by correct image binarization in advance. This is because GEI-based methods have been widely used in gait recognition, and they are based on the accumulated binarized image of human body in successive images. For that, the body geometric centers of successive images should be accurately aligned in order to obtain the correct movement information of human gait. If the segmented region of human body is not accurate, the calculated geometric center is not correct, either, which causes the extraction of incorrect movement information of human gait and consequent recognition error increases. In addition, the noise regions connected to the segmented human legs can causes the decrease of recognition accuracy. However, the accurate segmentation of human body including legs is difficult task requiring much processing time in the multiple and continuous images by visible light camera of surveillance environments due to the various environmental factors such as the variations of illuminations and shadows, etc. In addition, it is often the case that the leg parts of human (which are essential information for conventional gait recognition [78–85]) are not visible in our experimental images as shown in Figures 9–11 and 15.

However, we use the roughly detected region of body in a single image as shown in Figure 2d for recognition without the accurate segmentation of human body region and the alignment of body geometric center. It reduces the processing complexity and the performance of our system can be less affected by the detection accuracy of body regions. Even in the case that legs are not visible in the captured image, our method can correctly recognize human as shown in the 1st, 4th, 5th, 6th row images of Figure 15c and the 1st and 6th row images of Figure 15f.

As shown in Table 13, we compared the processing speed by our method with that by gait-based method [78]. Experimental platform is explained at the end of Section 5.1. As explained, the accurate segmentation of human body is important. However, experimental result showed that the segmentation performance based on background subtraction was bad with our experimental database due to the various factors of illumination variation and shadow, etc. Therefore, we adopted the deep learning-based segmentation method [86] for body segmentation, which was fine-tuned with our experimental database. As shown in Table 13, the processing speed per an image by our method is much faster than that by previous method.

Table 13. Comparison of processing time per an image by proposed and previous method (unit: ms).

Method	**Body Segmentation & Alignment**	**Matching Based on Radon Transform and PCA**	**Total**
Gait-based method [78]	752	145	897
Method	**Face & body detection**	**Matching based on two CNNs**	**Total**
Proposed method	98	327	425

In future, we are planning a study to improve recognition performance by automatically recreating the parts of the body region that cannot be seen in the images using a generative adversarial network (GAN). In addition, we plan to improve recognition performance by using super-resolution reconstruction to restore long-distance low-resolution images and make them into high-resolution images.

6. Conclusions

This paper proposed a multimodal human recognition method that uses both the face and body regions in indoor surveillance camera environments, and is based on deep CNNs (VGG Face-16 CNN and ResNet-50 CNN) by score-level fusion of Weighted Sum rule. Unlike previous methods, the proposed method recognizes the face and body regions in a single image separately and combines them to perform recognition in cases where the subject is approaching or moving further away from the camera, which occur frequently in an indoor surveillance camera environment (particularly hallways). In addition, whole body images of people are not used as input for a CNN. Instead, the face and body regions are separated and used as input for separate CNNs. Thus, the system can be used to recognize more detailed texture, color, and shape data for each region, and consequently, it can achieve better recognition accuracy than methods that use a whole body image as input for a single CNN. Unlike previous methods that focus only on cases where the entire body is included in the input images, the proposed method performs recognition on images where part of the body cannot be seen in the input images. To make impartial comparison experiments possible, we have publicly released [53] the VGG Face-16 and ResNet-50 CNN models which were trained in this study, along with the DFB-DB1 database which was custom made using two kinds of cameras to evaluate the performance of the proposed method in a variety of camera environments. In performance evaluations based on EER, ROC curves and CMC curves, it was confirmed that the proposed method (the EERs of 1.52% for DFB-DB1 and 0.58% for the ChokePoint dataset, and the GARs of rank1 of about 99.3% for DFB-DB1 and 99.95% for the ChokePoint dataset) is superior in comparison to face or body single modality-based recognition and other methods used in previous studies. However, FA and FR occurred in cases in which there was a big shape change between the enrolled images and the recognition image (particularly when part of the body could not be seen), as well as cases in which the image was captured at a long distance and had very poor resolution and cases in which there were large changes in the person's pose between images.

Author Contributions: J.H.K. and K.R.P. designed the face and body-based human recognition system based on two CNNs. S.W.C., N.R.B., and M.C.K. helped to experiments and analyzed results, and collecting databases.

Acknowledgments: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07041921), by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A1B03028417), and by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2018-2013-1-00684) supervised by the IITP (Institute for Information & communications Technology Promotion). Portions of the research in this paper use the ChokePoint collected by the National ICT Australia Ltd. (NICTA).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ahonen, T.; Hadid, A.; Pietikäinen, M. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 2037–2041. [CrossRef] [PubMed]
2. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A Unified Embedding for Face Recognition and Clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
3. Parkhi, O.M.; Vedaldi, A.; Zisserman, A. Deep Face Recognition. In Proceedings of the British Machine Vision Conference, Swansea, UK, 7–10 September 2015; pp. 1–12.
4. Nakajima, C.; Pontil, M.; Heisele, B.; Poggio, T. Full-body Person Recognition System. *Pattern Recognit.* **2003**, *36*, 1997–2006. [CrossRef]
5. Li, S.Z.; Lu, J. Face Recognition Using the Nearest Feature Line Method. *IEEE Trans. Neural Netw.* **1999**, *10*, 439–443. [CrossRef] [PubMed]
6. Turk, M.; Pentland, A. Eigenfaces for Recognition. *J. Cogn. Neurosci.* **1991**, *3*, 71–86. [CrossRef] [PubMed]
7. Etemad, K.; Chellappa, R. Discriminant Analysis for Recognition of Human Face Images. *J. Opt. Soc. Am.* **1997**, *14*, 1724–1733. [CrossRef]
8. Hong, H.G.; Lee, M.B.; Park, K.R. Convolutional Neural Network-Based Finger-Vein Recognition Using NIR Image Sensors. *Sensors* **2017**, *17*, 1297. [CrossRef] [PubMed]
9. Lee, M.B.; Hong, H.G.; Park, K.R. Noisy Ocular Recognition Based on Three Convolutional Neural Networks. *Sensors* **2017**, *17*, 2933.
10. Marcolin, F.; Vezzetti, E. Novel Descriptors for Geometrical 3D Face Analysis. *Multimed. Tools Appl.* **2017**, *76*, 13805–13834. [CrossRef]
11. Moos, S.; Marcolin, F.; Tornincasa, S.; Vezzetti, E.; Violante, M.G.; Fracastoro, G.; Speranza, D.; Padula, F. Cleft Lip Pathology Diagnosis and Foetal Landmark Extraction via 3D Geometrical Analysis. *Int. J. Interact. Des. Manuf.* **2017**, *11*, 1–18. [CrossRef]
12. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G. Emotion Recognition in Human-computer Interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80. [CrossRef]
13. Tsapatsoulis, N.; Doulamis, N.; Doulamis, A.; Kollias, S. Face Extraction from Non-uniform Background and Recognition in Compressed Domain. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Seattle, WA, USA, 15 May 1998; pp. 2701–2704.
14. Nguyen, D.T.; Hong, H.G.; Kim, K.W.; Park, K.R. Person Recognition System Based on a Combination of Body Images from Visible Light and Thermal Cameras. *Sensors* **2017**, *17*, 605. [CrossRef] [PubMed]

15. Kamgar-Parsi, B.; Lawson, W.; Kamgar-Parsi, B. Toward Development of a Face Recognition System for Watchlist Surveillance. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1925–1937. [CrossRef] [PubMed]
16. An, L.; Kafai, M.; Bhanu, B. Dynamic Bayesian Network for Unconstrained Face Recognition in Surveillance Camera Networks. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2013**, *3*, 155–164. [CrossRef]
17. Grgic, M.; Delac, K.; Grgic, S. SCface–Surveillance Cameras Face Database. *Multimed. Tools Appl.* **2011**, *51*, 863–879. [CrossRef]
18. Banerjee, S.; Das, S. Domain Adaptation with Soft-Margin Multiple Feature-Kernel Learning Beats Deep Learning for Surveillance Face Recognition. *arXiv* **2016**.
19. Taigman, Y.; Yang, M.; Ranzato, M.A.; Wolf, L. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1–8.
20. Antipov, G.; Berrani, S.-A.; Ruchaud, N.; Dugelay, J.-L. Learned vs. Hand-Crafted Features for Pedestrian Gender Recognition. In Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 1263–1266.
21. Layne, R.; Hospedales, T.M.; Gong, S. Towards Person Identification and Re-Identification with Attributes. In Proceedings of the European Conference on Computer Vision, Firenze, Italy, 7–13 October 2012; pp. 402–412.
22. Nguyen, D.T.; Park, K.R. Body-Based Gender Recognition Using Images from Visible and Thermal Cameras. *Sensors* **2016**, *16*, 156. [CrossRef] [PubMed]
23. Figueira, D.; Bazzani, L.; Minh, H.Q.; Cristani, M.; Bernardino, A.; Murino, V. Semi-Supervised Multi-Feature Learning for Person Re-Identification. In Proceedings of the 10th IEEE International Conference on Advanced Video and Signal Based Surveillance, Kraków, Poland, 27–30 August 2013; pp. 111–116.
24. Bak, S.; Corvee, E.; Brémond, F.; Thonnat, M. Person Re-Identification Using Spatial Covariance Regions of Human Body Parts. In Proceedings of the 7th IEEE International Conference on Advanced Video and Signal Based Surveillance, Boston, MA, USA, 29 August–1 September 2010; pp. 435–440.
25. Prosser, B.; Zheng, W.-S.; Gong, S.; Xiang, T. Person Re-Identification by Support Vector Ranking. In Proceedings of the British Machine Vision Conference, Aberystwyth, UK, 31 August–3 September 2010; pp. 1–11.
26. Chen, D.; Yuan, Z.; Chen, B.; Zheng, N. Similarity Learning with Spatial Constraints for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1268–1277.
27. Liao, S.; Hu, Y.; Zhu, X.; Li, S.Z. Person Re-Identification by Local Maximal Occurrence Representation and Metric Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2197–2206.
28. Li, W.; Zhao, R.; Xiao, T.; Wang, X. DeepReID: Deep Filter Pairing Neural Network for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 152–159.
29. Varior, R.R.; Haloi, M.; Wang, G. Gated Siamese Convolutional Neural Network Architecture for Human Re-Identification. In Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 791–808.
30. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Deep Metric Learning for Person Re-Identification. In Proceedings of the 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 34–39.
31. Shi, H.; Yang, Y.; Zhu, X.; Liao, S.; Lei, Z.; Zheng, W.; Li, S.Z. Embedding Deep Metric for Person Re-identification: A Study Against Large Variations. In Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 732–748.
32. Yang, Y.; Wen, L.; Lyu, S.; Li, S.Z. Unsupervised Learning of Multi-Level Descriptors for Person Re-Identification. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 4306–4312.
33. Wang, L.; Tan, T.; Ning, H.; Hu, W. Silhouette Analysis-Based Gait Recognition for Human Identification. *IEEE Trans. Pattern Anal. Mach. Intel.* **2003**, *25*, 1505–1518. [CrossRef]
34. Han, J.; Bhanu, B. Statistical Feature Fusion for Gait-Based Human Recognition. In Proceedings of the IEEE Conference and Computer Vision and Pattern Recognition, Washington, DC, USA, 27 June–2 July 2004; pp. II-842–II-847.

35. Zhou, X.; Bhanu, B. Integrating Face and Gait for Human Recognition at a Distance in Video. *IEEE Trans. Syst. Man Cybern. Part B-Cybern.* **2007**, *37*, 1119–1137. [CrossRef]
36. Zhou, X.; Bhanu, B. Feature Fusion of Side Face and Gait for Video-Based Human Identification. *Pattern Recognit.* **2008**, *41*, 778–795. [CrossRef]
37. Zhou, X.; Bhanu, B. Feature Fusion of Face and Gait for Human Recognition at a Distance in Video. In Proceedings of the 18th International Conference on Pattern Recognition, Hong Kong, China, 20–24 August 2006; pp. 529–532.
38. Zhou, X.; Bhanu, B.; Han, J. Human Recognition at a Distance in Video by Integrating Face Profile and Gait. In Proceedings of the Audio- and Video-based Biometric Person Authentication, Rye, NY, USA, 20–22 June 2005; pp. 533–543.
39. Kale, A.; RoyChowdhury, A.K.; Chellappa, R. Fusion of Gait and Face for Human Identification. In Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada, 17–21 May 2004; pp. V-901–V-904.
40. Gruber, I.; Hlaváč, M.; Železný, M.; Karpov, A. Facing Face Recognition with ResNet: Round One. In Proceedings of the International Conference on Interactive Collaborative Robotics, Hatfield, UK, 12–16 September 2017; pp. 67–74.
41. Martinez-Diaz, Y.; Mendez-Vazquez, H.; Lopez-Avila, L. Toward More Realistic Face Recognition Evaluation Protocols for the YouTube Faces Database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 526–534.
42. Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; Tian, Q. Person Re-Identification in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3346–3355.
43. Martinel, N.; Dunnhofer, M.; Foresti, G.L.; Micheloni, C. Person Re-Identification via Unsupervised Transfer of Learned Visual Representations. In Proceedings of the 11th International Conference on Distributed Smart Cameras, Stanford, CA, USA, 5–7 September 2017; pp. 151–156.
44. Shakhnarovich, G.; Lee, L.; Darrell, T. Integrated Face and Gait Recognition from Multiple Views. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, 8–14 December 2001; pp. I-439–I-446.
45. Shakhnarovich, G.; Darrell, T. On Probabilistic Combination of Face and Gait Cues for Identification. In Proceedings of the 5th IEEE Conference on Automatic Face and Gesture Recognition, Washington, DC, USA, 20–21 May 2002; pp. 169–174.
46. Guan, Y.; Wei, X.; Li, C.-T.; Marcialis, G.L.; Roli, F.; Tistarelli, M. Combining Gait and Face for Tackling the Elapsed Time Challenges. In Proceedings of the 6th IEEE Conference on Biometrics: Theory, Applications and Systems, Washington, DC, USA, 29 September–2 October 2013; pp. 1–8.
47. Hofmann, M.; Schmidt, S.M.; Rajagopalan, A.N.; Rigoll, G. Combined Face and Gait Recognition Using Alpha Matte Preprocessing. In Proceedings of the 5th IAPR International Conference on Biometrics, New Delhi, India, 29 March–1 April 2012; pp. 390–395.
48. Liu, Z.; Sarkar, S. Outdoor Recognition at a Distance by Fusing Gait and Face. *Image Vis. Comput.* **2007**, *25*, 817–832. [CrossRef]
49. Geng, X.; Wang, L.; Li, M.; Wu, Q.; Smith-Miles, K. Distance-Driven Fusion of Gait and Face for Human Identification in Video. In Proceedings of the Image and Vision Computing New Zealand, Hamilton, New Zealand, 5–7 December 2007; pp. 19–24.
50. Khamis, S.; Kuo, C.-H.; Singh, V.K.; Shet, V.D.; Davis, L.S. Joint Learning for Attribute-Consistent Person Re-Identification. In Proceedings of the European Conference on Computer Vision Workshops, Zurich, Switzerland, 6–7 September 2014; pp. 134–146.
51. Köstinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P.M.; Bischof, H. Large Scale Metric Learning from Equivalence Constraints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2288–2295.
52. Li, W.; Wang, X. Locally Aligned Feature Transforms across Views. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3594–3601.
53. Dongguk Face and Body Database (DFB-DB1). Available online: http://dm.dgu.edu/link.html (accessed on 16 June 2018).
54. Viola, P.; Jones, M.J. Robust Real-Time Face Detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154. [CrossRef]

55. Kazemi, V.; Sullivan, J. One Millisecond Face Alignment with an Ensemble of Regression Trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1867–1874.
56. OpenCV. Available online: https://opencv.org/ (accessed on 7 September 2018).
57. Kang, B.J.; Park, K.R. A Robust Eyelash Detection Based on Iris Focus Assessment. *Pattern Recognit. Lett.* **2007**, *28*, 1630–1639. [CrossRef]
58. Logitech BCC950 Camera. Available online: https://www.logitech.com/en-roeu/product/conferencecam-bcc950 (accessed on 25 January 2018).
59. Logitech C920 Camera. Available online: http://support.logitech.com/en_roeu/product/hd-pro-webcam-c920/specs (accessed on 25 January 2018).
60. ChokePoint Dataset. Available online: http://arma.sourceforge.net/chokepoint/ (accessed on 26 January 2018).
61. VGG Face-16 CNN Model. Available online: http://www.robots.ox.ac.uk/~vgg/software/vgg_face/ (accessed on 7 September 2018).
62. Huang, G.B.; Ramesh, M.; Berg, T.; Learned-Miller, E. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. In Proceedings of the Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition, Marseille, France, 17 October 2008; pp. 1–11.
63. Wolf, L.; Hassner, T.; Maoz, I. Face Recognition in Unconstrained Videos with Matched Background Similarity. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 529–534.
64. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–14.
65. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A Large-scale Hierarchical Image Database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
66. CS231n Convolutional Neural Networks for Visual Recognition. Available online: http://cs231n.github.io/convolutional-networks/#overview (accessed on 25 January 2018).
67. Convolutional Neural Network. Available online: https://en.wikipedia.org/wiki/Convolutional_neural_network (accessed on 25 January 2018).
68. Scherer, D.; Müller, A.C.; Behnke, S. Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition. In Proceedings of the 20th International Conference on Artificial Neural Networks, Thessaloniki, Greece, 15–18 September 2010; pp. 92–101.
69. ResNet-50 CNN Model. Available online: https://github.com/KaimingHe/deep-residual-networks (accessed on 7 September 2018).
70. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
71. Batch Norm Layer. Available online: http://caffe.berkeleyvision.org/tutorial/layers/batchnorm.html (accessed on 13 April 2018).
72. Bottou, L. Large-scale Machine Learning with Stochastic Gradients Descent. In Proceedings of the 19th International Conference on Computational Statistics, Paris, France, 22–27 August 2010; pp. 177–186.
73. Geforce GTX 1070. Available online: https://www.nvidia.com/en-us/geforce/products/10series/geforce-gtx-1070-ti/ (accessed on 27 May 2018).
74. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv* **2014**.
75. Visual Studio 2013. Available online: https://www.microsoft.com/en-us/search/result.aspx?q=visual+studio+2013 (accessed on 5 June 2018).
76. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification with Deep Convolutional Neural Networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
77. Receiver Operating Characteristic. Available online: https://en.wikipedia.org/wiki/Receiver_operating_characteristic (accessed on 20 February 2018).

78. Ali, H.; Dargham, J.; Ali, C.; Moung, E.G. Gait Recognition Using Gait Energy Image. *Int. J. Signal Process.* **2011**, *4*, 141–152.
79. Bouchrika, I.; Goffredo, M.; Carter, J.; Nixon, M. On Using Gait in Forensic Biometrics. *J. Forensic Sci.* **2011**, *56*, 882–889. [CrossRef] [PubMed]
80. Chen, J. Gait Correlation Analysis Based Human Identification. *Sci. World J.* **2014**, *2014*, 168275. [CrossRef] [PubMed]
81. Deshmukh, P.R.; Shelke, P.B. Gait Based Human Identification Approach. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **2016**, *6*, 495–498.
82. Guan, Y.; Li, C.-T. A Robust Speed-Invariant Gait Recognition System for Walker and Runner Identification. In Proceedings of the International Conference on Biometrics, Madrid, Spain, 4–7 June 2013; pp. 1–8.
83. Kusakunniran, W.; Wu, Q.; Zhang, J.; Li, H. Gait Recognition Across Various Walking Speeds Using Higher Order Shape Configuration Based on a Differential Composition Model. *IEEE Trans. Syst. Man Cybern.* **2012**, *42*, 1654–1668. [CrossRef] [PubMed]
84. Lv, Z.; Xing, X.; Wang, K.; Guan, D. Class Energy Image Analysis for Video Sensor-Based Gait Recognition: A Review. *Sensors* **2015**, *15*, 932–964. [CrossRef] [PubMed]
85. Wu, Z.; Huang, Y.; Wang, L.; Wang, X.; Tan, T.A. Comprehensive Study on Cross-View Gait Based Human Identification with Deep CNNs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 209–226. [CrossRef] [PubMed]
86. Arsalan, M.; Naqvi, R.A.; Kim, D.S.; Nguyen, P.H.; Owais, M.; Park, K.R. IrisDenseNet: Robust Iris Segmentation Using Densely Connected Fully Convolutional Networks in the Images by Visible Light and Near-Infrared Light Camera Sensors. *Sensors* **2018**, *18*, 1501. [CrossRef] [PubMed]